THIS WEEK

EDITORIALS

WORLD VIEW China can raise the stakes in global climate game **p.115**

ARCTIC Radar reveals how snowy ice cap is going thin on top **p.116**

IMAGING Nanotubes help peer through skull of mouse into brain **p.117**

Save the children

Infants and young people are being traumatized by armed conflict in their countries. Their resulting mental illnesses must be addressed, for the good of both the individuals and their society.

Battles in Ukraine, Gaza and Syria have appalled all who watch them from afar. The effects on the young provoke much of the horror. But many other armed conflicts are occurring, often far less visibly, in developing countries — and these are also home to the world's highest populations of children and young people.

Under-18s are described as requiring special protection in times of war in the United Nations Convention on the Rights of the Child, which celebrates its 25th anniversary this year. The convention, although lacking the teeth of enforcement, has provided a framework for discussions and planning that has spawned useful research. That research has begun to identify what 'special protection' really means — and the amount of time and resources it demands.

For a country to recover from war and rebuild a functional society, its young generation must be physically and mentally fit. In the past decade or so, humanitarian organizations have become increasingly aware of the prevalence of mental illness. This is particularly relevant for children and adolescents, because research has shown beyond doubt that prolonged and severe stress can damage the developing brain. Poor countries, often confronted with life-threatening epidemics of infectious disease, are too often unable to make mental illness a priority. But they surely need to embed in their health-care systems mental-health strategies for helping their traumatized youth.

Researchers, often supported by humanitarian organizations, have already undertaken scores of field studies in countries damaged by war or natural disasters. From Africa to Indonesia to the Balkans, they have tried to work out which interventions could help to mitigate or avert the mental damage caused by severe stress. Common interventions involve structured individual or group psychotherapy based in schools, for example, or family counselling.

As one might expect, the quality of societal support — an intact family, a trusted care-giver, a protective neighbourhood — has a major impact on whether an intervention will help. Still, many children emerge from trauma undamaged, even without an intervention. And an approach that works well in one context may even be harmful in another; for example, some displaced boys in Burundi responded negatively to a type of psychotherapy that had proved helpful in Indonesia.

There can be no single approach to limiting the mental damage inflicted by war. To be useful, interventions require intense engagement in the life and experience of each individual. For example, when working in Bosnia in the 1990s, a US psychiatrist discovered from conversations with one boy in his study group that, to get to school, the boy had to pass the tree from which he had witnessed his brother being hanged. It was helpful to bring this nightmare confrontation into their therapeutic sessions.

Worryingly, new scientific results are not getting through. Many popular therapeutic approaches — family counselling, for one — have not been rigorously tested in post-conflict contexts. And psychotherapy, known to be effective in post-traumatic stress disorder, is rarely practised, in part because of a lack of capacity to deliver it.

Humanitarian organizations, for all their importance, might not leave conflict zones with sound infrastructure. This underlines again the need for countries to develop their own scientific and medical capacities.

Immediate interventions in schools make sense, because rebuilding a society requires an educated next generation. But many more longitudinal studies are also needed to track traumatized children into adulthood, to see if and how the treatment they received helped them. ■

Future computing

Pushing the boundaries of current computing technologies will show the way to new ones.

What emerging technologies promise to displace conventional silicon chips? Future computers could run on graphene, perhaps, or the hidden powers of quantum physics or brain-like synaptic networks. Research on all these options and more is under way as it becomes clear that enhancement of silicon-chip technology is hitting serious practical obstacles: in manufacturing, connectivity and heat generation.

No emerging technology is likely to be a get-out-jail-free card. Amazing performance in one area is often accompanied by serious limits in another. Computing based on carbon nanotubes or graphene, for example, presents formidable challenges in reliable fabrication.

On page 147, information technologist Igor Markov argues that we should focus on the fundamental limits in computing, and use those to evaluate future possibilities. This approach has a rich history. Working out the maximum efficiency of steam engines, nineteenth-century physicists discovered thermodynamics. Modern information science was born in 1948 when Claude Shannon at Bell Labs considered what an ideal communication channel would look like.

Computations have limits: they take up space, time and energy. In 2000, IT researcher Seth Lloyd calculated the computing power of the ultimate laptop, which, by miraculous engineering advances, could harness all its energy for information processing (S. Lloyd *Nature* **406,** 1047–1054; 2000). This ultimate machine could perform 10^{51} operations per second, 40 orders of magnitude more than computers today. That represents 250 years of progress at current rates of improvement.

Markov's message is not to be overly optimistic or pessimistic about further progress. We should focus on the boundaries and push to see where they yield. ■

WORLD VIEW *A personal take on events*

China should aim for a total cap on emissions

A focus on carbon intensity alone will allow emissions to grow with the economy, argues **Qiang Wang**.

The big players in climate-change negotiations are starting to position themselves ahead of talks concerning a new global treaty in Paris next year.

Until now, the global deadlock on efforts to curb greenhouse-gas emissions has centred around the unwillingness of the United States to commit to a binding reduction target. This was shown most vividly by the nation's rejection of the 1997 Kyoto Protocol.

Many countries, China included, had little incentive to introduce policies to control carbon dioxide while the United States was not doing so. In June, the United States signalled a shift from that position when its Environmental Protection Agency (EPA) unveiled a new climate plan.

Using its authority under the Clean Air Act in lieu of congressional action, the EPA set a target to cut carbon pollution from power plants — the largest source of total US emissions — by 30% below 2005 levels by 2030.

Is the move a climate game changer? I believe that China will make some effort to react to the US plan. Exactly how is still unclear, but here is a suggestion: China, the world's biggest greenhouse-gas emitter, should upgrade its climate policy from reducing carbon intensity to setting a long-term cap on total emissions.

The difference is important. Carbon intensity is measured relative to gross domestic product, so while the economy is growing, so too can pollution. An absolute cap attempts to break that link: economic growth must not drive up carbon emissions.

CHINA MUST BE **CONVINCED** THAT THE LINK BETWEEN ECONOMIC GROWTH AND **EMISSIONS** CAN BE **BROKEN.**

In June, China's long-standing chief climate negotiator, Xie Zhenhua, gave the strongest signal yet that the country was considering such a switch. He told reporters at a meeting in Berlin that China was approaching a "peaking year" for its carbon emissions in the build-up to the Paris talks.

To agree on an emissions cap, China must be convinced that the link between economic growth and emissions can be broken. Here, there is another strong positive message from the United States. Nine states in the northeast of the country have started a cap-and-trade programme known as the Regional Greenhouse Gas Initiative, in which the government places a ceiling on carbon emissions and allows companies to buy and sell permits for those emissions. Since 2009, the states involved in the programme have cut their emissions by 18% on average, while their economies have grown by 9.2%. By comparison, emissions in the other 41 states fell by 4%, and their economies grew by 8.8%. Thus, the real challenge is not to set a cap for emissions, but to develop policies that make economic growth compatible with carbon reduction.

➲ NATURE.COM
Discuss this article online at:
go.nature.com/8tpsoo

If China does not set a carbon cap, then it could find it harder to continue to cut carbon intensity. With domestic coal demand in the United States expected to fall by 30% owing to the EPA rule, US coal firms — sitting on the largest recoverable reserves in the world — are pushing to increase exports to Asia, especially to China. Three new coal-export ports are being proposed for the Pacific coast, and are projected to ship up to 100 million tonnes of coal per year. The huge added supply to Asia will lead to cheaper coal and increased consumption. The European Union (EU) is a good example. Coal consumption has risen in the EU in recent years, and use of comparatively clean gas has fallen. This is partly because US coal exports to the EU sharply increased from 14 million tonnes in 2003 to 47 million tonnes in 2013.

It is unrealistic for China to switch immediately from cutting carbon intensity to a cap on emissions. A more rational and practical strategy is to make the transition in two steps.

First, China needs to obtain better data. Researchers must work out when Chinese emissions are likely to peak, assuming that the economy continues to grow as expected. This will provide a reliable baseline for any reduction target. It will require international scientific cooperation, because modelling for China must be informed by research results about the trajectory of emissions patterns in the EU, United States and other developed regions.

The peaking year is a complex issue and Chinese scientists and scholars differ greatly in their opinions of it. But the widely accepted view is that China's carbon output under the business-as-usual scenario will peak sometime after 2030.

Second, China needs to prioritize the use of 'bridging' fuel. It is no coincidence that the nine US states participating in the regional scheme have more nuclear energy and shale gas in their portfolios than most.

In 2011, nuclear energy accounted for less than 2% of China's electricity, but 12% of electricity globally and 21% in member countries of the Organisation for Economic Co-operation and Development. China's technically recoverable shale-gas resources are 31.6 trillion cubic metres, nearly double the United States' 18.8 trillion cubic metres. I advocate nuclear energy and shale gas as bridging fuels to a carbon-free future, if China can handle the safety and environmental concerns.

An absolute cap on China's emissions is in sight. But it will take political courage and practical changes to make it a reality. ■

Qiang Wang *is conjoint professor at the Xinjiang Institute of Ecology and Geography of the Chinese Academy of Sciences in Urumqi.*
e-mail: qiangwang7@gmail.com

The views expressed in this article are those of the author alone.

RESEARCH HIGHLIGHTS

Selections from the scientific literature

GENE EDITING

CRISPR corrects β-thalassaemia

A common genetic blood disorder has been corrected in cultured stem cells by using a cutting-edge genome-editing technique.

The disorder β-thalassaemia is characterized by reduced levels of haemoglobin due to mutations in the gene for β-globin (*HBB*). Yuet Kan and his colleagues at the University of California, San Francisco, created induced pluripotent stem cells using skin fibroblasts from a person with β-thalassaemia. They then used the CRISPR–Cas9 gene-editing technique to correct the unwanted mutation precisely, without affecting other genes. After differentiation in culture into precursors of red blood cells, the modified cells showed higher expression of *HBB* than unmodified cells.

Transplantation of such corrected cells back into the original patient could one day provide a cure for β-thalassaemia, say the authors.

Genome Res. **http://doi.org/t3v (2014)**

ASTRONOMY

Another super-Earth found

A ‘super-Earth’ planet — an extrasolar planet larger than Earth but smaller than Neptune — has been detected in the habitable zone of a star called Gliese 832.

Robert Wittenmyer at the University of New South Wales in Sydney, Australia, and his colleagues used data from various telescopes to detect a planet with a mass of 5.4 Earths in orbit around this star. Although the planet is in the habitable zone — the region around a star in which it is thought that life could potentially exist — its large size suggests that it may have a thick atmosphere. This might make it more like a ‘super-Venus’, with a dense atmosphere leading to high surface temperatures that would render it inhospitable.

Despite this, the presence of this potentially rocky inner planet, as well as a previously discovered outer giant planet, makes the Gliese 832 system a rare miniature version of our Solar System, the authors suggest.

Astrophys. J. 791, **114 (2014)**

CHRIS LINDER/UW

POLAR SCIENCE

Arctic snowpack thins

As Arctic sea ice has shrunk and thinned, so has the snowpack blanketing it.

Melinda Webster at the University of Washington in Seattle and her colleagues studied data on spring snow depth gathered between 2009 and 2013 by radar surveys conducted from the air and verified with surface measurements (pictured). They compared these to information collected between 1954 and 1991 by Soviet ice stations. The error bars are large, but between the older and the current surveys, snow thickness had decreased by some 37% in the western Arctic and by 56% in the Beaufort and Chukchi seas.

As sea ice starts forming later each autumn, there is less time for snow to accumulate before winter sets in, the authors say.

J. Geophys. Res. Oceans **http://doi.org/t3q (2014)**

CHEMISTRY

Cleaner, greener ammonia

A method of producing ammonia could yield a greener route to nitrogen-based fertilizers.

Ammonia is currently synthesized by combining nitrogen and hydrogen under high pressures and temperatures in a reaction called the Haber–Bosch process. Making the hydrogen consumes around 5% of the world’s natural-gas production, and releases large amounts of carbon dioxide.

Stuart Licht at George Washington University in Washington DC and his colleagues applied a voltage to steam and air (the source of nitrogen) bubbling through molten hydroxide containing catalytic nanoparticles of iron oxide. This produced ammonia from nitrogen and water directly by electrolysis. The nanoparticles clump together over time, slowing the reaction, and moderate temperatures and pressures are still needed. However, if the process can be scaled up, it could be less energy-intensive than the current industrial method.

Science 345, **637–640 (2014)**

MICROBIOLOGY

Resistance genes mapped

Researchers have pinpointed mutations encoding antibiotic resistance in bacteria that cause pneumonia, borrowing a technique more often used to hunt for gene variations linked to common human diseases.

Streptococcus pneumoniae is a leading killer of children under five worldwide. The bacterium is prone to develop antibiotic resistance, but pinning down the mutations responsible has proved difficult.

A team led by Stephen Bentley and Julian Parkhill, at the Wellcome Trust Sanger Institute in Hinxton, UK, analysed the genomes of 3,701 samples of *S. pneumoniae* collected from carriers in a refugee camp in Thailand and from patients in Massachusetts clinics.

The authors searched for regions of the genome that differed between bacteria resistant to β-lactam antibiotics (such as penicillin) and those still susceptible to them. They found 301 DNA variations in 51 regions linked to drug resistance, including novel genes as well as those involved in building the cell wall, the target of the β-lactams.

PLoS Genet. 10, **e1004547 (2014)**

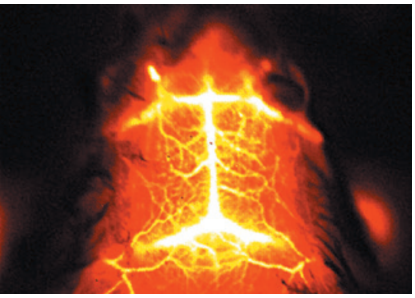
IMAGING

Seeing through a mouse skull

Glowing nanotubes have allowed researchers to peer through a mouse's skull and examine its living brain in real time.

Calvin Kuo and Hongjie Dai of Stanford University in California and their colleagues injected fluorescent molecules based on carbon nanotubes into the tails of mice. The nanotubes were then carried around in the animals' bloodstreams and when lasers were shone onto the rodents' skulls, the molecules gave off near-infrared light (**pictured**) that was visible through the bone. This allowed the researchers to image blood moving through the brain to a depth of more than 2 millimetres and to detect obstructed arteries. However, the method might not be usable in humans because of our thicker skulls.

Nature Photon. **http://doi.org/t2z (2014)**



DAI LAB AND KUO LAB, STANFORD UNIV.

SEISMOLOGY

From earthquakes to icequakes

Big earthquakes on land can trigger small distant 'icequakes' in the Antarctic ice sheet.

At magnitude 8.8, the 2010 Maule earthquake in Chile was the largest quake in the Southern Hemisphere for half a century. Zhigang Peng at the Georgia Institute of Technology in Atlanta and his colleagues hunted for traces of it at seismic stations across Antarctica.

They discovered high-frequency shaking representing small icequakes, with waves of tremors appearing in the kilometre-thick ice sheet that covers the frozen continent. These seemed to be triggered by the lower-frequency rumble stemming from the Chilean event, and represent the first evidence of links between quakes in the solid earth and in the cryosphere.

Nature Geosci. **http://dx.doi.org/10.1038/ngeo2212 (2014)**

STEM CELLS

Fresh growth from elderly cells

Human skin cells can be reprogrammed into neural cells that form synapses with neurons in severed spinal cords in rats.

A team led by Paul Lu and Mark Tuszynski at the University of California San Diego in La Jolla took skin fibroblasts from an 86-year-old man, converted them in culture into induced pluripotent stem cells (iPS cells) and then into neural stem cells, and grafted these cells into two-week-old immunodeficient rats whose spinal cords were damaged at the neck. Three months later, the stem cells had grown into neurons that projected axons along the whole length of the rat spinal cord, even extending into the brain. Unlike similar experiments with neurons derived from embryonic stem cells, these iPS-cell-derived neurons did not restore movement in the rats' limbs, perhaps as a result of scar tissue that formed at the injury site.

Neuron **http://doi.org/t36 (2014)**

SOCIAL SELECTION *Popular articles on social media*

Clash over the Kardashians of science

Here's a novel approach for getting an article noticed: put 'Kardashian' in the title. A paper that compared Twitter-using researchers to the celebrity Kim Kardashian incited a backlash on social media.

Neil Hall, a genomics researcher at the University of Liverpool, UK, introduced a metric called the Kardashian Index, or K value. This is calculated by dividing a researcher's number of Twitter followers by the number of scientific citations he or she has. The K value supposedly identifies scientists whose visibility exceeds their contributions — somewhat like a certain socialite, Hall suggests. The article was intended as satire, but not everyone was amused. "This paper suggests only highly cited scientists deserve a large Twitter following, & everyone else should shut up," tweeted Katie Mack, an astrophysicist at the University of Melbourne in Australia.

Genome Biol. 15, **424 (2014)**

Based on data from altmetric.com. Altmetric is supported by Macmillan Science and Education, which owns Nature Publishing Group.

➲ NATURE.COM
For more on popular papers:
go.nature.com/hqeqwb

MICROBIOLOGY

Ecosystems afloat in asphalt

Water droplets suspended in the world's largest tar 'lake' are teeming with diverse ecosystems of bacteria and methane-producing microorganisms, despite the inhospitable living conditions.

Droplets just a few microlitres in volume that were isolated from Pitch Lake (**pictured**), a huge tar pit on the island of Trinidad, contain a menagerie of bacteria and archaea, report Rainer Meckenstock at the Helmholtz Zentrum in Munich, Germany, and his colleagues.

CARLOS BARRIA/REUTERS/CORBIS

They used DNA sequencing to reveal that multiple species work together to break down the oil surrounding the water droplets, which are thought to originate deep underground.

These microhabitats could be an unrecognized factor in the biodegradation of large volumes of oil, the authors suggest.

Science 345, **673–676 (2014)**

For a longer story on this research, see go.nature.com/odleal

➲ NATURE.COM
For the latest research published by *Nature* visit:
www.nature.com/latestresearch

SEVEN DAYS *The news in brief*

ESA/ROSETTA/MPS FOR OSIRIS TEAM MPS/UPD/LAM/IAA/SSO/INTA/UPM/DASP/IDA

Rosetta's rendezvous

The European Space Agency's comet-chasing spacecraft Rosetta arrived at its destination on 6 August after a ten-year journey. Performing the last of a set of ten manoeuvres, Rosetta entered the same orbit around the Sun as its target, 67P/Churyumov–Gerasimenko, to become the first spacecraft to rendezvous with a comet. The probe will study the body before attempting to place a lander, Philae, on its surface in November. Rosetta will continue to follow and measure the comet as it swings around the Sun in August 2015. See go.nature.com/oqzeaa for more.

POLICY

US–Africa summit

US President Barack Obama backed the idea of a Global Alliance for Climate-Smart Agriculture at the US–Africa Leaders Summit in Washington DC on 4–6 August. The partnership would bring together governments, industry and non-governmental organizations to help boost African agriculture while keeping farming-related greenhouse-gas emissions in check. The alliance is slated for launch on 23 September. The summit also saw Sweden pledge US$1 billion to Obama's Power Africa initiative to double people's access to electricity in sub-Saharan Africa.

Emissions lawsuit

A coalition of environmental groups is mounting a legal challenge to force the US Environmental Protection Agency (EPA) to regulate greenhouse-gas emissions from aircraft. On 5 August, the groups, led by the Center for Biological Diversity in Tucson, Arizona, filed a formal notice of intent to sue. They argue that the EPA should impose limits on aviation emissions under the Clean Air Act — the law that regulates carbon dioxide produced by power plants and vehicles.

Licence battle

A coalition of more than 50 research institutions, funders and open-access publishers signed a letter dated 7 August protesting against a new set of licences governing open-access articles (see go.nature.com/agficr). The licences will limit the legal reuse of research articles and data that are supposed to be freely available to the public, the coalition argues. The Association of Scientific, Technical and Medical Publishers, a trade group headquartered in Oxford, UK, drew up the disputed licences. The letter calls for the Creative Commons licences to be used as the global standard for open research output.

Timber law

Illegal wood products are still entering European markets one year after the start of a law to prevent trade in illicit timber. A survey conducted by the WWF, the international environmental group, found that only 11 European Union countries have adopted national legislation and robust penalties. The worst culprits include Hungary and Spain, said the WWF on 6 August. The survey echoes an assessment from the European Commission that found similar failings. Illegal logging is a leading cause of deforestation in tropical forests.

GMO green light

The US Department of Agriculture says that new varieties of genetically engineered maize (corn) and soya beans will not become plant 'pests' to other crops. The agency's final environmental assessment, published on 6 August, paves the way for approval of the first plants engineered to be resistant to the herbicide 2,4-D. Cotton and soya-bean plants engineered to resist the herbicide dicamba also passed the assessment. The US Environmental Protection Agency is still reviewing the herbicides to be used on the crops.

EVENTS

Ballute flight

A balloon–parachute hybrid designed by NASA to slow down spacecraft entering the thin atmosphere of Mars is ready for use, the agency announced on 8 August. The 'ballute' was one of two re-entry devices tested on a 28 June flight over the Pacific Ocean near Hawaii. The second, a supersonic parachute, tore on deployment. NASA intends to test a redesigned version on two more test flights next year.

Troublesome book

More than 130 leading population geneticists have condemned a book that argues that genetic variation between human populations could underlie global economic, political and social differences. The book, *A Troublesome Inheritance* (Penguin, 2014), by science journalist Nicholas Wade, uses "incomplete and inaccurate explanations" of research to support arguments about differences among human societies, the geneticists say in a 10 August letter to *The New York Times*. Wade's book was published in May. See go.nature.com/ktvblx for more.

Ebola emergency

The World Health Organization (WHO) declared the West African Ebola outbreak a public-health emergency of international concern on 8 August, just before deaths reached more than 1,000. The outbreak is still concentrated in Sierra Leone, Guinea and Liberia, but Nigeria is also reporting

NSB/ZOJ/NICK STERN/WENN.COM/NEWSCOM

cases. On 11 August, the WHO convened a meeting of experts to discuss the ethics of using experimental medicines that have not yet been tested in humans. In a statement, the panel concluded that "it is ethical to offer unproven interventions with as yet unknown efficacy and adverse events, as potential treatment or prevention". Two Americans have already received an experimental antibody, made by Mapp Biopharmaceutical of San Diego, California, and further doses of the scarce drug are to be shipped to West Africa. See go.nature.com/9u6sic for more.

FACILITIES

Polar power cut

Research has stalled at the British Antarctic Survey's Halley Research Station (**pictured**) in Antarctica after a power failure, the organization said in a statement on 6 August. Six days later, the station's staff reported that a coolant leak from a main pipe had occurred on 30 July, leading to generators overheating and shutting down. Some power and heating has been restored, but "all science, apart from meteorological observations essential for weather forecasting, has been stopped". Disrupted work includes ozone monitoring, meteorology for climate science, and studies of the upper atmosphere used for forecasting space weather. See go.nature.com/cjtrpt for more.

Suez upgrade

Egypt has announced a US$4-billion construction project to add an extra channel to the Suez canal to allow more ships to pass along this vital trade route from the Red Sea to the Mediterranean. The 5-year project involves digging or dredging along 72 kilometres of the canal's 163-kilometre length, say officials at the Suez Canal Authority. On 5 August, Egypt's president Abdel Fattah el-Sisi said that he hopes to see the new waterway opened in one year from now.

PEOPLE

Biologist dies

J. Woodland Hastings, who helped to found the study of circadian rhythms, died on 6 August at his home in Lexington, Massachusetts. He was 87. Hastings studied bacterial bioluminescence, including its day–night rhythms, at Harvard University in Cambridge, Massachusetts. His research also provided early evidence of bacterial communication and quorum sensing — a system that lets species detect and respond to stimuli according to their population density.

Scripps leader

The Scripps Research Institute in La Jolla, California, announced on 11 August that it has appointed one of its molecular biologists, James Paulson, as acting president. The institute — which has a US$21-million budget deficit — is still looking for a long-term leader after the resignation of Michael Marletta in July. Marletta had stepped down after his plan for a $600-million merger between Scripps and the University of Southern California in Los Angeles triggered a faculty revolt (see go.nature.com/cvozom). He will remain a staff member at Scripps.

RESEARCH

Corrupt ivory

All legal sales of ivory should be stopped for at least ten years because corruption is ruining attempts to save African elephants, according to a paper published on 7 August by the Wildlife Conservation Society (see E. L. Bennett *Conserv. Biol.* http://doi.org/t5v; 2014). The society, based in New York, finds that corruption among government officials in charge of legal ivory markets is aggravating conservation problems. It points out that six of the eight countries identified as the worst offenders in ivory trafficking are in the bottom half of a league table of honest governance and public services drawn up by Transparency International in Berlin. See go.nature.com/x8jvew for more.

COMING UP

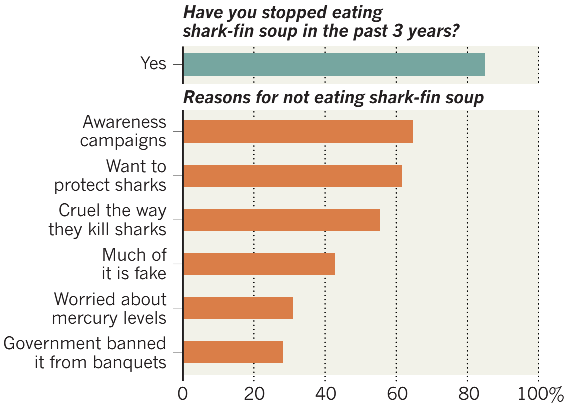
16–21 AUGUST
The World Weather Open Science Conference in Montreal, Canada, discusses how to improve seasonal predictions. Topics also include the dynamics and predictability of weather systems such as clouds and tropical cyclones.
go.nature.com/7uskg7

TREND WATCH

An online survey of 1,568 Chinese consumers suggests that demand for shark-fin soup is falling, according to a 4 August report by WildAid, a non-governmental organization in San Francisco, California. The report also polled shark-fin vendors in Guangzhou, China, who reported declined sales and reduced prices, and fishermen in Indonesia who said that prices had dropped. WildAid says that awareness campaigns and the desire to protect sharks are major factors in changing attitudes.

LOSS OF APPETITE FOR SHARK-FIN SOUP

A survey finds that Chinese consumer demand for the delicacy is falling (as are traders' prices).



SOURCE: WILDAID

➲ NATURE.COM
For daily news updates see:
www.nature.com/news

DAVID MCNEW/GETTY

Unprecedented drought in California has substantially degraded aquatic habitats.

CLIMATE

Native ecosystems blitzed by drought

California's current water crisis offers a preview of what climate change will bring.

BY ALEXANDRA WITZE

Peter Moyle has seen a lot in five decades of roaming California's streams and rivers and gathering data on the fish that live in them. But last month he saw something new: tributaries of the Navarro River, which rises in vineyards before snaking through a redwood forest to the Pacific, had dried up completely.

"They looked in July like they normally look in September or October, at the end of the dry season," says Moyle, a fish biologist at the University of California, Davis.

Blame the drought. The Navarro and its hard-pressed inhabitants are just one example of stresses facing a parched state. From the towering Sierra Nevada mountains — where the snowpack this May was only 18% of the average — to the broad Sacramento–San Joaquin river delta, the record-setting drought is reshaping California's ecosystems.

It is also giving researchers a glimpse of the future. California has always had an extreme hydrological cycle, with parching droughts interrupted by drenching Pacific storms (see 'Extreme hydrology'). But scientists say that the current drought — now in its third year — holds lessons for what to expect 50 years from now.

"The west has always gone through this, but we'll be going through it at perhaps a more rapid cycle," says Mark Schwartz, a plant ecologist and director of the John Muir Institute of the Environment at the University of California, Davis. He and others are discussing the drought's ecological consequences at the annual meeting of the Ecological Society of America, which runs from 10 to 15 August in Sacramento, California. He says that the state's plant and animal species are at risk in part because California ecosystems are already highly modified and vulnerable to a variety of stresses.

Many of the state's 129 species of native inland fish, including several types of salmon, are listed by federal or state agencies under various levels of endangerment. "We're starting from a pretty low spot," says Moyle. He hopes to use the current drought to explore where native fish have the best chances of surviving.

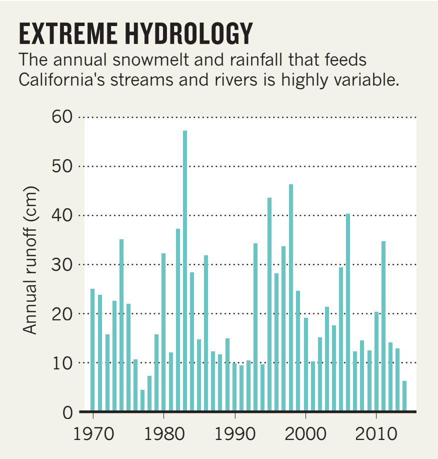
That could be in dammed streams such as Putah Creek near the Davis campus, where water flow can be controlled to optimize native fish survival. Another focus might be on spring-fed streams such as those that flow down from volcanic terrain in northernmost California and can survive drought much longer than snow-fed streams.

In the late 1970s, Moyle discovered that native fish in the Monterey Bay watershed recolonized their streams relatively quickly after a two-year drought. But today's streams face greater ecological pressures, such as more dams and more non-native species competing for habitat.

SPACE INVADERS

Other challenges arise in the delta where the Sacramento and San Joaquin rivers meet, northeast of San Francisco. An invasive saltwater clam (*Potamocorbula amurensis*) has taken advantage of warming river waters and moved several kilometres upriver, says Janet Thompson, an aquatic ecologist with the US Geological Survey (USGS) in Menlo Park, California.

Potamocorbula out-competes a freshwater clam (*Corbicula fluminea*), and accumulates about four times as much of the element selenium from agricultural run-off and refineries as its freshwater cousin does. When endangered sturgeon feed on *Potamocorbula*, the fish consume much more selenium than is optimal. "That's the biggest shift that we've seen that's of environmental concern," says Thompson. ▶



SOURCE: USGS

"These are the kinds of things that can have a lasting effect on a predator species."

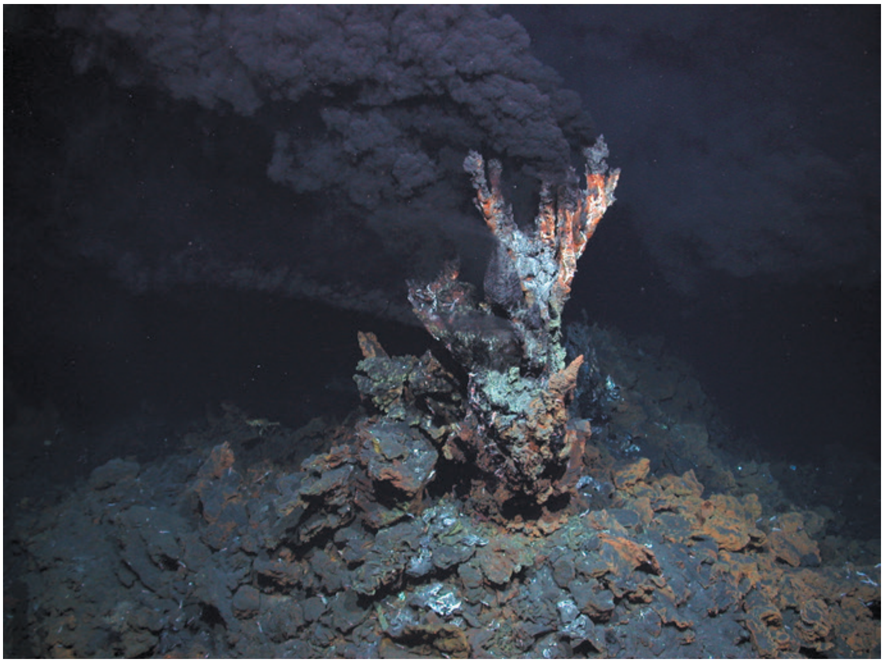
Teasing out the drought's effects on terrestrial animals is tougher. Researchers have documented drops in various California bird populations this year, such as mallard ducks (*Anas platyrynchos*) and tricolor blackbirds (*Agelaius tricolor*). But many other factors — especially habitat loss — also come into play, so it becomes hard to isolate the effects of drought.

The drought's effects on larger animals such as bears are also uncertain. Anecdotal reports suggest that more bears than usual are showing up closer to people this year, says Jason Holley, a wildlife biologist at the California Department of Fish and Wildlife in Rancho Cordova. Within the space of six weeks this spring, four black bears appeared along the Sacramento River corridor, much farther out of the mountains than normal. "Those sorts of calls definitely pique your interest," says Holley, who thinks that dry conditions in the mountains might be pushing bears closer to populated areas.

The longest-lasting effect could be on California's forests, including its iconic giant sequoias. The drought has handed forest ecologists an unplanned experiment, says Phillip van Mantgem, a forestry expert at the USGS in Arcata, California, who is speaking at the Sacramento meeting.

Researchers are gathering data to examine whether thinning of plots in the forest, in part to reduce fire risk, might help trees do better under drought. Tests may also help to reveal the main mechanisms by which drought kills different tree species, whether by interrupting the flow of water within the tree or by starving it. "I'm really curious to see how this turns out," van Mantgem says.

There should be plenty of time to gather data. Climatologists expect an El Niño weather pattern to form in the Pacific this year, which usually brings more rain and snow to parts of California (see *Nature* **508,** 20–21; 2014). But the pending El Niño looks to be weaker than first expected, and may not have much, if any, influence on ending the drought. Chances are that the state will remain dry well into 2015. ■



MARUM, UNIV. BREMEN

Marine communities living near mining targets such as hydrothermal vent fields might be at risk.

MARINE SCIENCE

Health check for deep-sea mining

European project evaluates risks to delicate ecosystems.

BY KATIA MOSKVITCH

As commercial plans to exploit mineral resources on deep-ocean beds gather pace, marine researchers are increasingly concerned about the damage such projects might cause to the sensitive and little-understood ecosystems that thrive there. Now, scientists are taking to the sea as part of a three-year, €12-million (US$16-million) project designed to address these concerns and to develop a set of guidelines for industry.

The latest research expedition of the Managing Impacts of Deep-sea Resource Exploitation (MIDAS) programme returned to France earlier this month after exploring the Lucky Strike region of the Mid-Atlantic Ridge near the Azores islands. There, a research team began investigating whether plumes of particles that might arise from future mining operations near hot hydrothermal vents — often rich sources of metals — could affect the creatures that live there, such as deep-sea mussels.

"The goal of our experiment is to test the effects of sulphide particle deposits on the structure — composition, density, biomass, diversity — of the dominant hydrothermal fauna of the Lucky Strike vent field," says Jozée Sarrazin, a deep-sea ecologist at the French Research Institute for Exploitation of the Sea (IFREMER) in Plouzané, France, who is leading the expedition. "It should help us to propose management strategies to help protect the unique fauna associated with high-temperature emissions on the sea floor."

Resources such as polymetallic sulphides, manganese nodules, cobalt-rich ferromanganese crusts, methane hydrates and rare-earth elements exist in large quantities around deep-sea hydrothermal vents, having escaped from the molten crust below. The idea of mining them was first mooted in the 1960s, but only now, with land sources declining and demand rising, is it being seriously explored.

Although no mining projects are yet under way, Nautilus Minerals of Toronto, Canada, has received a green light from the government of Papua New Guinea to mine about 50 kilometres offshore in the Bismarck Sea, at a depth of 1.6 kilometres. Other concessions have been awarded in the eastern Pacific Ocean. Nautilus would use sea-floor trawlers to cut or scoop up the deposits, which are then pumped up to a support ship.

The effects of such mining are cause for concern. The operations may "severely damage" the sensitive biological communities that live near under-sea mountains, hydrothermal vents and mineral-rich nodules on the sea floor, says David Santillo, a marine biologist and senior scientist at Greenpeace Research Laboratories at the University of Exeter, UK. As well as the physical destruction of habitats, he adds, this type of mining could smother deep-sea species with suspended plumes of sediment. Species could also be disturbed by noise, light pollution and exposure to toxic metals and other chemicals released by the mining.

The severity of such effects depends on several factors, including the nature of the exploited resource and the method of extraction, says oceanographer Cindy Van Dover, director of the Duke University Marine Laboratory in Beaufort, North Carolina. But her biggest concern is the general lack of knowledge about sea-floor processes and the cumulative effects of multiple mining events. “If we get the environmental management wrong, we are unlikely to be able to fix our mistake,” she says.

The MIDAS project, which began in November, is receiving €9 million from the European Union, and includes representatives from industry and non-governmental organizations. “We will try to identify the best ways to monitor before, during and after mining to determine the total impact and recovery of the ecosystems,” says Philip Weaver, managing director of Seascape Consultants in Romsey, UK, which is coordinating MIDAS.

Cruises to conduct experiments and sampling at depth form a core part of the project’s work. The IFREMER cruise, on the research vessel *Pourquoi Pas?*, was the first stage of a two-year experiment to test the effects of sulphide plumes. The research team weighed mussels found around hydrothermal vents at a depth of 1.7 kilometres and assessed their general health. Next year, they will return and mimic the effects of particle plumes on the mussels, monitoring their reactions — for instance, death, migration or increased numbers — with temperature sensors and cameras. The results of the tests will then be studied on shore.

A second MIDAS study is currently simulating potential effects on marine life in the shallow waters of Portman Bay off the coast of southeastern Spain. An onshore mining facility dumped waste into the waters there for three decades, and the researchers want to assess how the waste affected the underwater fauna. “We want to see how metal-loaded plumes behave — how far they spread, how long it takes for them to settle and so on,” says marine geoscientist Miquel Canals Artigas of the University of Barcelona in Spain, who is leading the expedition.

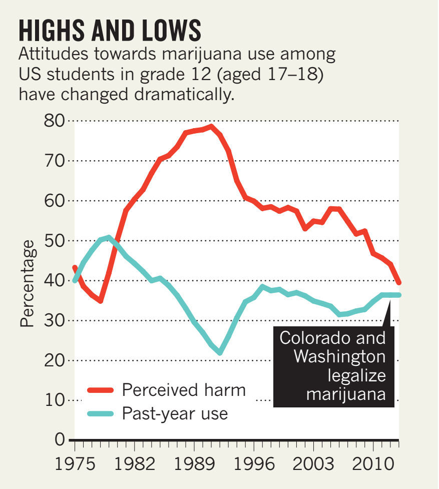
MIDAS will submit its report to the European Commission in November 2016. ■

NEUROSCIENCE

Teen drug use gets supersize study

US government programme will examine 10,000 adolescents to document effects on developing brains.

BY SARA REARDON



SOURCE: UNIV. MICHIGAN, 2013 ‘MONITORING THE FUTURE’ STUDY

When the states of Colorado and Washington voted to legalize marijuana in 2012, the abrupt and unprecedented policy switch sent the US National Institute on Drug Abuse (NIDA) into what its director Nora Volkow describes as “red alarm”. Although marijuana remained illegal for people under the age of 21, the drug’s increased availability and growing public acceptance suggested that teenagers might be more likely to try it (see ‘Highs and lows’). Almost nothing is known about whether or how marijuana affects the developing adolescent brain, especially when used with alcohol and other drugs.

The new laws, along with advances in brain-imaging technology, convinced Volkow to accelerate the launch of an ambitious effort to follow 10,000 US adolescents for ten years in an attempt to determine whether marijuana, alcohol and nicotine use are associated with changes in brain function and behaviour.

At a likely cost of more than US$300 million, it will be the largest longitudinal brain-imaging study of adolescents yet. Researchers are eager to study a poorly understood period of human development — but some question whether it is possible to design a programme that will provide useful information about the effects of drugs.

“It’s definitely an idea that’s overdue,” says Deanna Barch, a psychologist at Washington University in St. Louis, Missouri. “The downside is it’s a lot of eggs in one basket.”

The exact design of the programme is still in flux. In May, NIDA held a planning meeting with the National Institute on Alcohol Abuse and Alcoholism, the National Cancer Institute and the National Institute of Child Health and Development (NICHD), which will help to fund the project. The partners decided to recruit participants at the age of ten. Roughly every two years, researchers will image the children’s brains, perform psychiatric and cognitive tests, and examine factors such as genetics and environmental exposures.

To enlist enough participants likely to use drugs, the study will recruit largely from high-risk groups, such as children of low socio-economic status or those whose parents use drugs. Volkow says that the group plans to seek input from colleagues at November’s Society for Neuroscience meeting in Washington DC before recruiting researchers for the programme.

Hugh Garavan, a psychmjmeist at the University of Vermont in Burlington, says that there is much to recommend such an analysis. ▶

MORE ONLINE

TOP STORY

ASHLEY COOPER PICS/ALAMY

Canada–US oil pipeline could significantly increase emissions go.nature.com/jcgw9n

MORE NEWS

- ● Multiple genes linked to post-traumatic stress disorder go.nature.com/anfirh
- ● Anxiety drug helps boost survival of certain fish species go.nature.com/vbkxqg
- ● Origami robot folds itself in 4 minutes go.nature.com/swqnz7

He and his colleagues reported in July — on the basis of a study of 692 European teenagers — that they had identified brain structures and activity patterns that could predict with around 70% accuracy which 14-year-olds would become binge drinkers by the age of 16 (R. Whelan *et al. Nature* http://doi.org/t5t; 2014).

A larger study such as NIDA's is the obvious next step, he says, because it will help researchers to account for the myriad environmental and genetic factors that influence development. "I can anticipate this becoming a landmark study," he adds.

Even if the larger human study does not reveal new information about the effects of drugs, it will certainly shed light on the normally developing brain, says Lisa Freund, a developmental psychologist at the NICHD. Volkow says that NIDA plans to make all trial data available to researchers, which could inspire further studies.

But recruiting and retaining so many participants will be challenging, requiring researchers to win the trust of children and their families and to ensure that participants are not burdened by the drug tests and brain scans. The scientists will also need to be flexible in the face of rapidly improving brain-imaging techniques, says Terry Jernigan, a cognitive scientist at the University of California, San Diego. Whatever technology is chosen is likely to be obsolete by the study's end, she cautions; upgrading technology haphazardly could make it difficult to compare data across years or study sites. Researchers could use both new and old imaging methods during transition periods, but that could quickly drive up costs.

Others question whether enough is known about the developing brain to identify the mechanisms underlying certain specific conditions. Some studies have linked adolescent use of marijuana to psychosis and to onset of schizophrenia in those at risk, but it is unclear whether this is the case — and if so, whether the drug is a trigger or the teenagers are self-medicating. Because researchers cannot control the drug's timing and dosage, it will be hard to resolve that question in relation to what is happening in the brains of participants who develop psychosis.

"It's almost impossible to establish direct causality" in such cases, Volkow says. Yet she hopes that the NIDA study's huge sample size will reveal broader clues, such as differences in brain structure between drug users and non-users. The institute is considering running a parallel brain-imaging study on non-human primates to help to address this.

B. J. Casey, a psychologist at Cornell University in New York City, hopes that NIDA will address the concerns raised by scientists. "Once you start something like this, it's hard to stop even if the outcomes aren't telling, because people are so invested," she says. ■

A worker sprays insecticide in Haiti to fight mosquitoes that carry chikungunya and other diseases.

HECTOR RETAMAL/AFP/GETTY

INFECTIOUS DISEASE

US assesses virus of the Caribbean

Researchers warn that a change of mosquito host could accelerate spread of chikungunya across the Americas.

BY ALESZU BAJAK

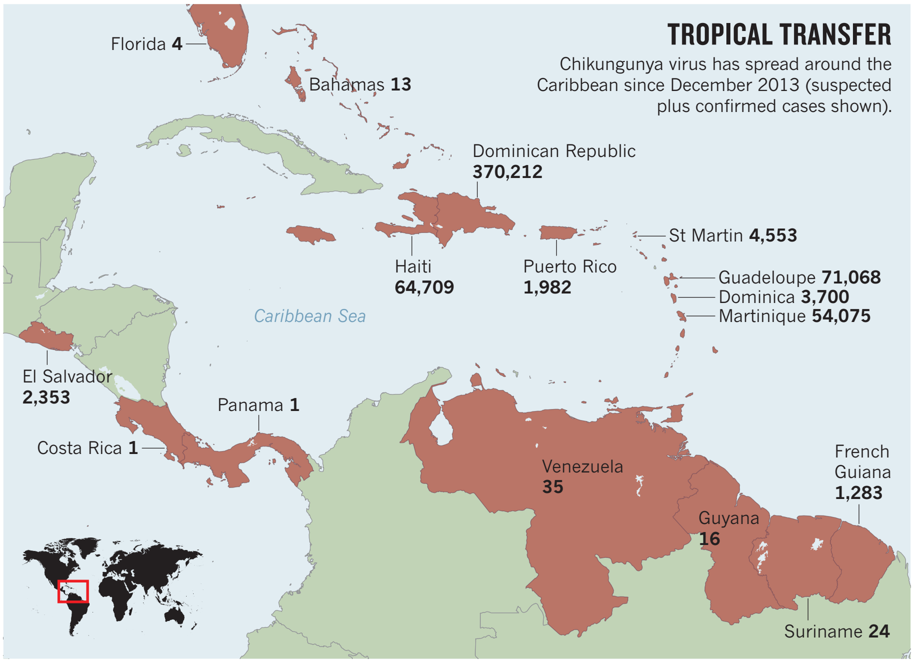
In the past few months, passengers at North American airports have been warned that travel to the Caribbean might result in an unwanted souvenir. The first outbreak of chikungunya virus in the Western Hemisphere began in the French part of the Caribbean island of St Martin in December and has spread rapidly around the region, infecting more than 500,000 people.

Since then, at least 480 travellers have returned to the United States with the mosquito-borne disease, raising concerns that an insect biting one of those people would spark a US chikungunya outbreak. Yet so far, only four locally acquired cases have been confirmed in the country, all in southern Florida. The virus has gained more of a foothold in Central and South America: authorities have confirmed 174 cases of locally transmitted disease in El Salvador, Panama, Costa Rica, Venezuela and the Guianas (see 'Tropical transfer').

For now, the Caribbean strain of chikungunya does not seem likely to expand into the rest of the Western Hemisphere, mostly because it is spread by the tropical mosquito *Aedes aegypti*. However, several major chikungunya outbreaks have been fuelled by a specific mutation of the virus that makes it more suited to transmission by a different species of mosquito — a scenario analysed by Carrie Manore, a mathematical epidemiologist at Tulane University in New Orleans, Louisiana, and her colleagues. They report that genetic changes in the virus could propel chikungunya deep into North and South America (C. A. Manore *et al. J. Theor. Biol.* **356,** 174–191; 2014). The insect that could cause the damage is the Asian tiger mosquito (*Aedes albopictus*), which has been expanding worldwide for the past two decades and taking diseases such as chikungunya and dengue with it (see *Nature* **489,** 187–188; 2012).

Chikungunya was first detected in the 1950s in East Africa. It causes fever, severe joint pain and, in rare cases, death. Most people recover within a week, but painful arthritic symptoms can linger for months.

The Caribbean is fertile ground for the



TROPICAL TRANSFER

Chikungunya virus has spread around the Caribbean since December 2013 (suspected plus confirmed cases shown).

SOURCE: PAHO/WHO

spread of the disease, no matter which mosquito is spreading it. In temperate regions, winter weather kills *A. aegypti* mosquitoes and acts as a natural brake on the spread of the diseases they carry. But in the Caribbean, *A. aegypti* can survive year-round, and serves as an outstanding host vector for diseases, says Sylvain Aldighieri, a physician at the Pan American Health Organization in Washington DC who has helped to track the current outbreak.

Native to Africa, *A. aegypti* had spread throughout the warmer zones of the Western Hemisphere by the seventeenth century. It is found across the southern United States, and has penetrated as far north as Virginia. In mainland South America, says Aldighieri, it can be found in every country except Chile. Still, the Caribbean might be the only place in the hemisphere with the right density of mosquitoes and travelling people to enable a chikungunya outbreak.

Scientists are worried about the rapid expansion of the Asian tiger mosquito, which is more aggressive than *A. aegypti* and more efficient at transmitting chikungunya. During a 2005 chikungunya outbreak on the island of Réunion, east of Madagascar, *A. albopictus* suddenly became a more efficient vector because a genetic mutation in the virus allowed it to reproduce better in the mosquito's midgut and to be transmitted more easily. The same mutation arose independently in the virus on the Indian Ocean island of Mayotte in 2006, and again in 2007 when it appeared in Madagascar.

If a similar mutated virus strain were introduced to the Western Hemisphere — or if, as in the past, the Caribbean strain were to mutate — chikungunya would become a much larger public-health concern for the Americas.

In North America, the Asian tiger mosquito is found in 32 states, from New York to Texas, and has been spotted in California, New Mexico and Arizona. Data for the Southern Hemisphere are less prevalent and not so reliable, says epidemiologist David Morens of the US National Institute of Allergy and Infectious Diseases in Bethesda, Maryland. However, the species is known to be widespread in Latin America.

Manore and her colleagues used a mathematical model to assess the danger of a chikungunya outbreak spread by the Asian tiger mosquito. It considers rates of susceptibility, infectiousness and immunity in both humans and mosquitoes to predict how an outbreak might evolve over time. The researchers find that the relative risk and severity of an outbreak depend on the virus–mosquito combination, with the highest risk coming from the Asian tiger mosquito carrying the Réunion mutant strain of chikungunya.

"I'd be concerned about areas that have both *Aedes albopictus* and *Aedes aegypti*," Manore says. In these regions — where the virus could most easily jump to the more aggressive species — there is an urgent need for more mosquito trapping and studies of how the virus and the insect interact, she says.

Even within countries, different subgroups of the same mosquito species sometimes transmit the same viral strain differently, which makes comprehensive sampling a pressing matter. And tracking something as pervasive and hard to pin down as a mosquito is no simple task, says Erin Staples, a medical epidemiologist at the US Centers for Disease Control and Prevention in Atlanta, Georgia. "Right now, chikungunya is definitely one of our concerns," she says. "How well will it be transmitted? We don't know." ■

SCIENTISTS AND THE SOCIAL NETWORK

Giant academic social networks have taken off to a degree that no one expected even a few years ago. **A *Nature* survey explores why.**

BY RICHARD VAN NOORDEN

In 2011, Emmanuel Nnaemeka Nnadi needed help to sequence some drug-resistant fungal pathogens. A PhD student studying microbiology in Nigeria, he did not have the expertise and equipment he needed. So he turned to ResearchGate, a free social-networking site for academics, and fired off a few e-mails. When he got a reply from Italian geneticist Orazio Romeo, an international collaboration was born. Over the past three years, the two scientists have worked together on fungal infections in Africa, with Nnadi, now at Plateau State University in Bokkos, shipping his samples to Romeo at the University of Messina for analysis. “It has been a fruitful relationship,” says Nnadi — and they have never even met.

Ijad Madisch, a Berlin-based former physician and virologist, tells this story as just one example of the successes of ResearchGate, which he founded with two friends six years ago. Essentially a scholarly version of Facebook or LinkedIn, the site gives members a place to create profile pages, share papers, track views and downloads, and discuss research. Nnadi has uploaded all his papers to the site, for instance, and Romeo uses it to keep in touch with hundreds of scientists, some of whom helped him to assemble his first fungal genome.

More than 4.5 million researchers have signed up for ResearchGate, and another 10,000 arrive every day, says Madisch. That is a pittance compared with Facebook’s 1.3 billion active users, but astonishing for a network that only researchers can join. And Madisch has grand goals for the site: he hopes that it will become a key venue for scientists wanting to engage in collaborative discussion, peer review papers, share negative results that might never otherwise be published, and even upload raw data sets. “With ResearchGate we’re changing science in a way that’s not entirely foreseeable,” he says, telling investors and the media that his aim for the site is to win a Nobel prize.

The company now employs 120 people, and last June it announced that it had secured US$35 million from investors including the world’s richest individual, Bill Gates — cash that came on top of two earlier rounds of undisclosed investment. “It was really a head-scratcher when we saw that,” says Leslie Yuan, who heads a team working on networking and innovation software for scientists at the University of California, San Francisco. “We thought — who are these guys? How are they getting so much money?”

> “WE’RE CHANGING SCIENCE IN A WAY THAT’S NOT ENTIRELY FORESEEABLE.”

Yuan is not the only one who has been taken aback. A few years ago, the idea that millions of scholars would rush to join one giant academic social network seemed dead in the water. The list of failed efforts to launch a ‘Facebook for science’ included Scientist Solutions, SciLinks, Epernicus, 2collab and Nature Network (run by the company that publishes *Nature*). Some observers speculated that this was because scientists were wary of sharing data, papers and comments online — or if they did want to share, they would prefer do it on their own terms, rather than through a privately owned site.

But it seems that those earlier efforts were ahead of their time — or maybe they were simply doing it wrong. Today, ResearchGate is just one of several academic social networks going viral. San Francisco-based competitor Academia.edu says that it has 11 million users. “The goal of the company is to rebuild science publishing from the ground up,” declares chief executive Richard Price, who studied philosophy at the University of Oxford, UK, before he founded Academia.edu in 2008, and has already raised $17.7 million from venture capitalists. A third site, London-based Mendeley, claims 3.1 million members. It was originally launched as software for managing and storing documents, but it encourages private and public social networking. The firm was snapped up in 2013 by Amsterdam-based publishing giant Elsevier for a reported £45 million (US$76 million).

WINNING FORMULA

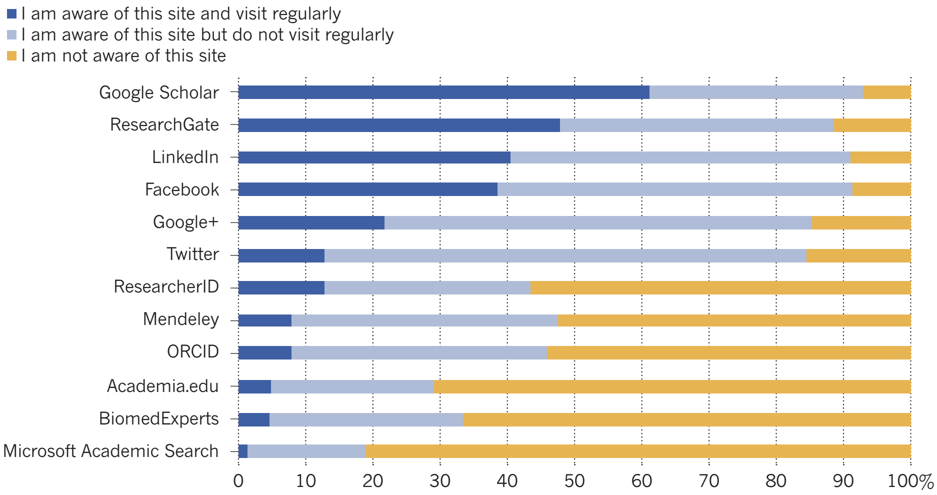
Despite the excitement and investment, it is far from clear how much of the activity on these sites involves productive engagement, and how much is just passing curiosity — or a desire to access papers shared by other users that they might otherwise have to pay for. “I’ve met basically no academics in my field with a favourable view of ResearchGate,” says Daniel MacArthur, a geneticist at Massachusetts General Hospital in Boston.

In an effort to get past the hype and explore what is really happening, *Nature* e-mailed tens of thousands of researchers in May to ask how they use social networks and other popular profile-hosting or search services, and received more than 3,500 responses from 95 different countries.

The results confirm that ResearchGate is

REMARKABLE REACH

More than 3,000 scientists and engineers (below) told Nature about their awareness of various giant social networks and research-profiling sites. Just under half said that they visit ResearchGate regularly. Another 480 respondents in the humanities, arts and social sciences were less keen on ResearchGate — see charts at go.nature.com/fjvxxt.



certainly well-known (see 'Remarkable reach', and full results online at go.nature.com/jvx7pl). More than 88% of scientists and engineers said that they were aware of it — slightly more than had heard of Google+ and Twitter — with little difference between countries. Just under half said that they visit regularly, putting the site second only to Google Scholar, and ahead of Facebook and LinkedIn. Almost 29% of regular visitors had signed up for a profile on ResearchGate in the past year.

This does not surprise Billie Swalla, an evolutionary biologist and director of the University of Washington's Friday Harbor Laboratories. Swalla says that she and most of her colleagues are on ResearchGate, where she finds the latest relevant papers much more easily than by following marine-biology journals. "They do send you a lot of spam," she says, "but in the past few months, I've found that every important paper I thought I should read has come through ResearchGate." Swalla admits to comparing herself to others using the site's 'RG Score' — its metric of social engagement. "I think it taps into some basic human instinct," she adds.

TACTICAL BREAKDOWN

Some irritated scientists say that the site taps into human instincts only too well — by regularly sending out automated e-mails that profess to come from colleagues active on the site, thus luring others to join on false pretences. (Indeed, 35% of regular ResearchGate users in *Nature*'s survey said that they joined the site because they received an e-mail.) Lars Arvestad, a computer scientist at Stockholm University, is fed up with the tactic. "I think it is a disgraceful kind of marketing and I am choosing not to use their service because of that," he says. Some of the apparent profiles on the site are not owned by real people, but are created automatically — and incompletely — by scraping details of people's affiliations, publication records and PDFs, if available, from around the web. That annoys researchers who do not want to be on the site, and who feel that the pages misrepresent them — especially when they discover that ResearchGate will not take down the pages when asked. Madisch is unruffled by these complaints. The pages are marked for what they are, and are not counted among the site's real users, he says, adding: "We changed many things based on the feedback we got. But the criticism is relatively small, relative to the number of people who like the service."

Academia.edu seems less well-known than ResearchGate: only 29% of scientists in the survey were aware of it and just 5% visited regularly. But it has its fans — among them climate scientist Hans von Storch, director of the Institute for Coastal Research in Geesthacht, Germany, who uses the site to share not only his papers, but also his interviews, book reviews and lectures. Price points out that Academia.edu has much higher web traffic than ResearchGate overall, perhaps because — unlike its rival — it is open to anyone to join. And for the 480 social science, arts and humanities researchers included in *Nature*'s survey, usage of the two sites was more closely matched.

➲ NATURE.COM
For an interactive graphic and more on profile-managing, see: go.nature.com/fjvxxt

> "WE HAVE TO BUILD BETTER FILTER SYSTEMS TO EXPLAIN WHAT RESEARCH YOU CAN TRUST."

High numbers by themselves do not mean much, says Jan Reichelt, a co-founder of Mendeley (which scored 48% awareness and 8% regular visitors among scientists in *Nature*'s survey). "We've moved away from mentioning 'start-up vanity metrics' as the key number," he says. "It doesn't tell you about the quality of interaction."

To get a rough measure of that quality, *Nature* asked a subset of the most active respondents what they actually do on the sites they visit regularly (see 'Idle, browse or chat?'). The most-selected activity on both ResearchGate and Academia.edu was simply maintaining a profile in case someone wanted to get in touch — suggesting that many researchers regard their profiles as a way to boost their professional presence online. After that, the most popular options involved posting content related to work, discovering related peers, tracking metrics and finding recommended research papers.

"These are tools that people are using to raise their profiles and become more discoverable, not community tools of social interaction," argues Deni Auclair, a lead analyst for Outsell, a media, information and technology consulting firm in Burlingame, California. By comparison, Twitter, although used regularly by only 13% of scientists in *Nature*'s survey, is much more interactive: half of the Twitterati said that they use it to follow discussions on research-related issues, and 40% said that it is a medium for "commenting on research that is relevant to my field" (compared with 15% on ResearchGate).

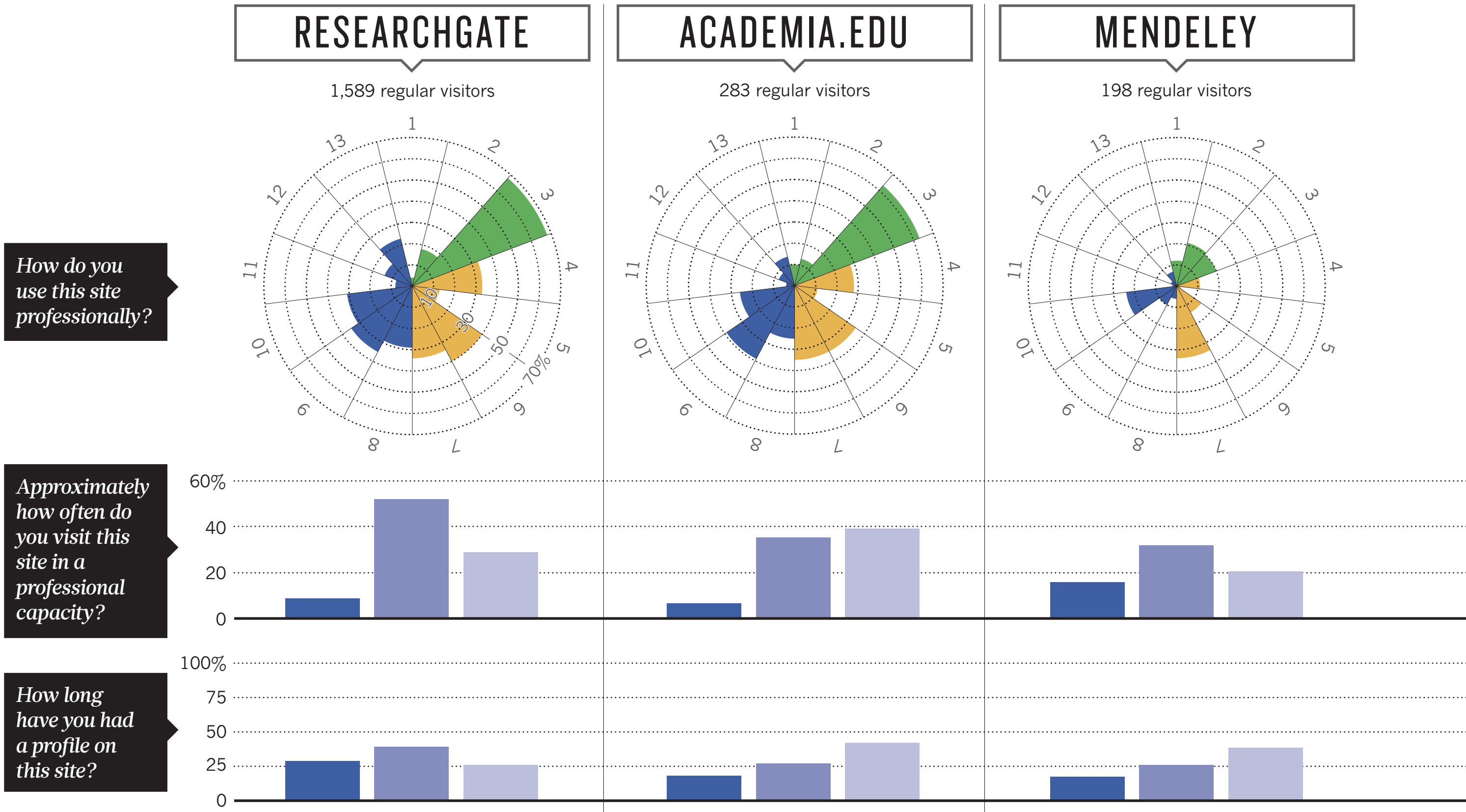
PAPERS PLEASE

Laura Warman, an ecologist at the Institute of Pacific Islands Forestry in Hilo, Hawaii, echoes the views of many when she says that she has uploaded papers on Academia.edu to keep track of how often, where and when they are downloaded. "I find it especially intriguing that my most downloaded paper is not my most cited work," she says. "To put it bluntly, I have no idea if these sites have any impact whatsoever on my career — I tend to doubt they do — but I enjoy knowing that my work is being discussed."

Price says that 3 million papers have been uploaded to Academia.edu, and Madisch says that 14 million are accessible through ResearchGate (although he will not say how many of those have been automatically scraped from freely accessible places elsewhere). An unpublished study conducted by computer scientists Madian Khabsa at Pennsylvania State University in University Park and Mike Thelwall at the University of Wolverhampton, UK, suggests that by August this year, the full texts of around one-quarter of all molecular-biology papers published in 2012 were available from ResearchGate. That said, these days papers are easily found on many sites: a study conducted

IDLE, BROWSE OR CHAT?

Nature asked a subset of regular visitors to social networks how they used the sites professionally. (Each person was asked to tick all activities that applied.) The results suggest that Facebook is not widely used professionally; that researchers on Twitter are very active and social; and that many users of Academia.edu and ResearchGate signed up in case someone wants to contact them — but are not chatty themselves. Full results are available at go.nature.com/jvx7pl.



"Mainly, a source of stress every time an e-mail pops in showing that my colleagues/competitors are publishing more than myself."
AGE 35–44, PROFESSOR, BRAZIL

"It is the only useful community website for research purposes."
AGE 55–64, PROFESSOR, HUNGARY

"Has led to invites to referee papers/external assessments."
AGE 45–54, POSTDOCTORAL FELLOW, UNITED KINGDOM

"I have been able to post old papers which otherwise would be inaccessible to people."
AGE 55–64, PROFESSOR, UNITED STATES

"Primarily still a reference manager for me. The social component is less important."
AGE 35–44, RESEARCH SCIENTIST, CANADA

"Fairly useful as a documents clearinghouse for lab group."
AGE 25–34, POSTDOCTORAL FELLOW, UNITED STATES

for the European Commission last year found that 18% of biology papers published in 2008–11 were open access from the start, and said that 57% could be read for free in some form, somewhere on the Internet, by April 2013 (see *Nature* **500,** 386–387; 2013).

Publishers are worried that the sites could become public troves of illegally uploaded content. In late 2013, Elsevier sent 3,000 notices to Academia.edu and other sites under the US Digital Millennium Copyright Act (DMCA), demanding that they take down papers for which the publisher owned copyright. Academia.edu passed each notice on to its users — a decision that triggered a public outcry. One researcher who received a take-down request did not want to be named, but told *Nature*: "I hardly know any scientists who don't violate copyright laws. We just fly below the radar and hope that the publishers don't notice."

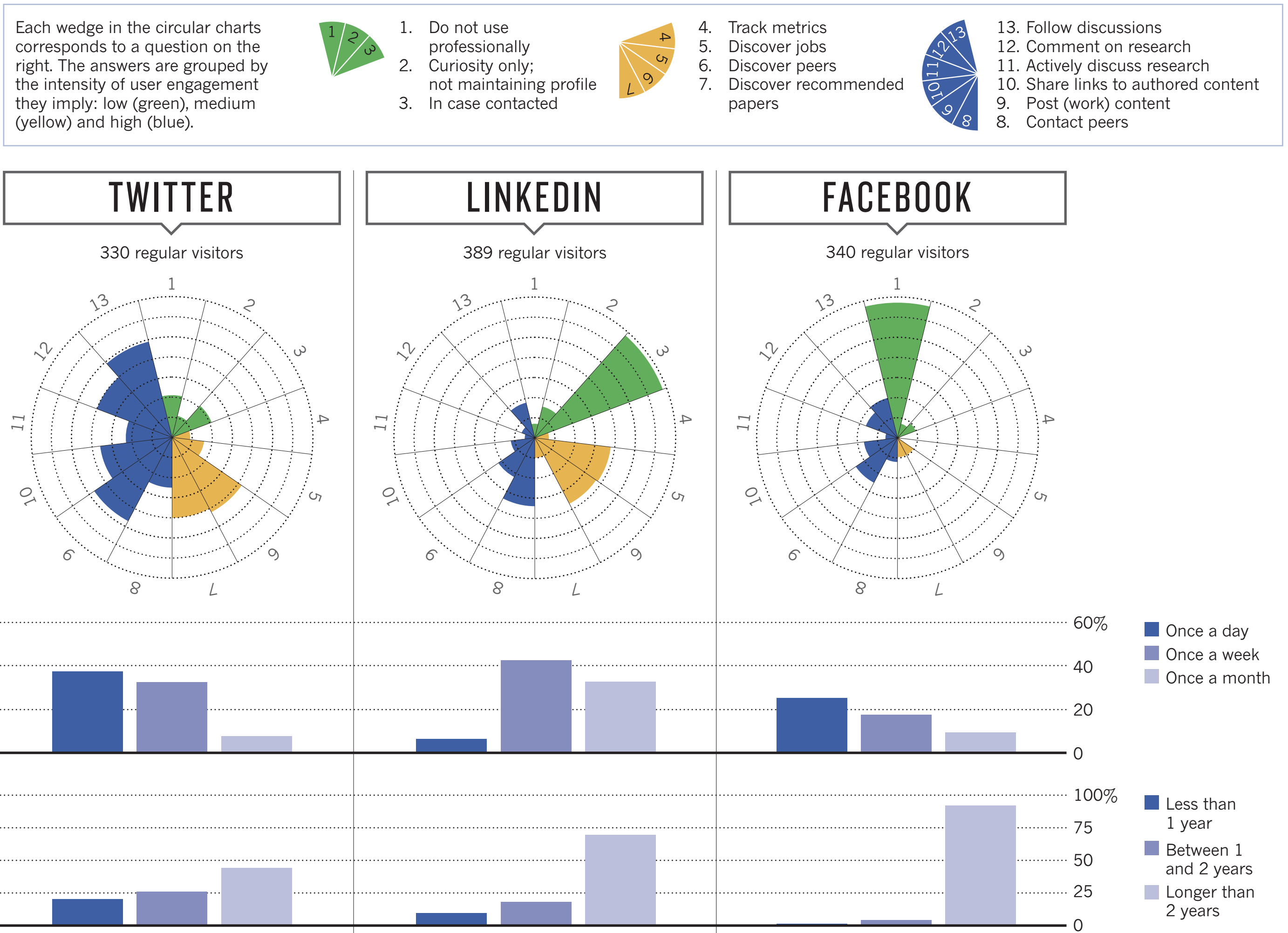
These concerns are not unique to large social networks, says Price; the same issue surrounds content posted in universities' online repositories (to which Elsevier also sent some DMCA notices last year). "This is really part of the wider battle where academics want to share their papers freely online, whereas publishers want to keep content behind a paywall to monetize it," he says, noting the nuance that many publishers allow researchers to upload the final accepted version of a manuscript, but not the final PDF. He has seen fewer take-down notices this year.

OPEN INTENTIONS

Giant social networks could also disrupt the research landscape by capturing other public content. In March this year, ResearchGate launched a feature called Open Review, encouraging users to post in-depth critiques of existing publications. Madisch says that members have now contributed more than 10,000 such reviews. "I believe that this is just the tip of the iceberg," he says. He wants users to upload raw data sets too — including, perhaps, negative results that might otherwise never be published — and says that 700 are appearing on the site each day.

At Academia.edu, Price is planning to launch a post-publication peer-review feature as well. "We have to build better filter systems to explain what research you can trust," he says.

Few would argue with these goals, but many wonder why researchers would deposit their data sets and reviews on these new social networks, rather than elsewhere online — on their own websites, for example, in university repositories, or on dedicated data-storage sites such as Dryad or figshare (see *Nature* **500,** 243–245; 2013 — figshare is funded by *Nature*'s parent company, Macmillan Publishers). To Madisch,



Each wedge in the circular charts corresponds to a question on the right. The answers are grouped by the intensity of user engagement they imply: low (green), medium (yellow) and high (blue).

1. Do not use professionally
2. Curiosity only; not maintaining profile
3. In case contacted

4. Track metrics
5. Discover jobs
6. Discover peers
7. Discover recommended papers

13. Follow discussions
12. Comment on research
11. Actively discuss research
10. Share links to authored content
9. Post (work) content
8. Contact peers

"Extremely useful in conference settings."
AGE 35–44, RESEARCH SCIENTIST, UNITED STATES

"Great way to keep up-to-date on what is happening NOW in the research community."
AGE 45–54, HEAD OF ACADEMIC DEPARTMENT, UNITED STATES

"Mainly useful for job hunting."
AGE 25–34, PHD STUDENT, UNITED STATES

"It is too much like Facebook — fluffy forwards and such that are not scientific or related to professionalism."
AGE 45–54, ASSOCIATE PROFESSOR, UNITED STATES

"Facebook has zero credibility in my professional life."
AGE 35–44, STAFF SCIENTIST, UNITED STATES

"The (invitation-only) groups for professional astronomers and pulsar astronomers have become vibrant discussion fora."
35-44, RESEARCH SCIENTISTS, UNITED STATES

the answer lies with the social sites' burgeoning communities of users — the famed 'network' effect. "If you post on ResearchGate, you are reaching the people who matter," he says. But Titus Brown, a computational scientist at Michigan State University in East Lansing, is concerned about the sites' business plans as they seek to survive. "What worries me is that at some point ResearchGate will use their information to make a profit in ways that we are uncomfortable with — or they will be bought by someone who will do that," he says.

Madisch says that ResearchGate will not sell its user data, and that it already makes some money by running job adverts (as does Academia.edu). In the future, he hopes to add a marketplace for laboratory services and products, connecting companies and corporate researchers to academics (28% of the network's users are from the corporate world, he says). Price talks about providing institutional analytics to universities as well. But analysts including Auclair argue that the sites have limited earning potential, because they are targeted at a much narrower demographic than Facebook or Twitter. "What's most likely is the networks that have critical mass get acquired and those that don't will die," she says (although Madisch says that being bought out "would be a personal failure").

Mendeley's acquisition by Elsevier last year left the site better placed to become a global platform for research collaboration, says Reichelt, because it intersects with other Elsevier products such as the Scopus database of research articles. Much of the collaboration done using Mendeley is private, but the firm does allow other computer programs to automatically pull out useful anonymized public information — such as which papers are viewed most by which researchers. Neither Academia.edu nor ResearchGate yet offer this service, although Madisch says that he is developing it.

"I think at some point there will be one winner in this race," says Madisch. Or — as *Nature*'s survey suggests is already happening — different disciplines might favour different sites. Some analysts argue that despite their millions of users, massive social academic networking sites have not yet proven their essential worth. "They are nice-to-have tools, not need-to-have," says Auclair. But Price says that the networks are on the front line of a trend that cannot be ignored. "We saw the changes in the market, and we could see that academics wanted to share openly. The tide is starting to turn in our direction." ■

Richard Van Noorden *is a senior reporter for* Nature *in London.*

MEGAMIX/GETTY

Don't blame the mothers

Careless discussion of epigenetic research on how early life affects health across generations could harm women, warn **Sarah S. Richardson** and colleagues.

From folk medicine to popular culture, there is an abiding fascination with how the experiences of pregnant women imprint on their descendants. The latest wave in this discussion flows from studies of epigenetics — analyses of heritable changes to DNA that affect gene activity but not nucleotide sequence. Such DNA modification has been implicated in a child's future risk of obesity, diseases such as diabetes, and poor response to stress.

Headlines in the press reveal how these findings are often simplified to focus on the maternal impact: 'Mother's diet during pregnancy alters baby's DNA' (BBC), 'Grandma's Experiences Leave a Mark on Your Genes' (*Discover*), and 'Pregnant 9/11 survivors transmitted trauma to their children' (*The Guardian*). Factors such as the paternal contribution, family life and social environment receive less attention.

Questions about the long shadow of the uterine environment are part of a burgeoning field known as developmental origins of health and disease (DOHaD)[1]. For example, one study revealed[2] that 45% of children born to women with type 2 diabetes develop diabetes by their mid-twenties, compared with 9% of children whose mothers developed diabetes after pregnancy.

DOHaD would ideally guide policies that support parents and children, but exaggerations and over-simplifications are making scapegoats of mothers, and could even increase surveillance and regulation of pregnant women. As academics working in DOHaD and cultural studies of science, we are concerned. We urge researchers, press officers and journalists to consider the ramifications of irresponsible discussion.

ALARMING PRECEDENTS

There is a long history of society blaming mothers for the ill health of their children. Preliminary evidence of fetal harm has led to regulatory over-reach. First recognized in the 1970s, fetal alcohol syndrome (FAS) is a collection of physical and mental problems in children of women who drink heavily during pregnancy. In 1981, the US Surgeon General advised that no level of alcohol consumption was safe for pregnant women. Drinking during pregnancy was stigmatized and even criminalized. Bars and restaurants were required to display warnings that drinking ▶

causes birth defects. Many moderate drinkers stopped consuming alcohol during pregnancy, but rates of FAS did not fall[3].

Although those who drink heavily during pregnancy can endanger their children, the risks of moderate drinking were overstated by policy-makers — a point recently reaffirmed by the Danish National Birth Cohort study, which did not find adverse effects in children whose mothers drank moderately during pregnancy[4]. Nonetheless, warnings about alcohol during pregnancy made in inappropriate contexts still cause pregnant women to suffer social condemnation and to agonize over an occasional sip.

In the 1980s and 1990s, surging use of crack cocaine (a smokable form of the drug) in the United States led to media hysteria around 'crack babies' — those who had been exposed to cocaine in the womb. Pregnant women who took drugs lost social benefits, had their children taken away and were even sent to prison. More than 400 pregnant women, mostly African American, have been prosecuted for endangering their fetuses in this way. Exposed infants were stigmatized as a biologically doomed underclass. Today, fetal exposure to crack or cocaine is considered no more harmful than exposure to tobacco or alcohol[5], but criminal prosecution of pregnant women who take such drugs continues.

Previous generations found other ways to blame women. As late as the 1970s, 'refrigerator mothers' (a disparaging term for a parent lacking emotional warmth) were faulted for their children's autism. Until the nineteenth century, medical texts attributed birth deformities, mental defects and criminal tendencies to the mother's diet and nerves, and to the company she kept during pregnancy.

Although it does not yet go to the same extremes, public reaction to DOHaD research today resembles that of the past in disturbing ways. A mother's individual influence over a vulnerable fetus is emphasized; the role of societal factors is not. And studies now extend beyond substance use, to include all aspects of daily life.

CONTEXT IS KEY

A 2013 story on the health-information website WebMD demonstrates the sort of responsible reporting that we would like to see more of (see go.nature.com/p2krhs). The story reported findings of a four-fold increased risk of bipolar disorder in adult offspring if a mother had influenza during pregnancy[6], but it emphasized that the overall risk observed was small and that bipolar disorder is treatable. It stated that the study considered only one of many possible risk factors and did not establish cause and effect. Furthermore, the headline did not lead with the scary number.

Much less context was given in coverage of a 2012 paper[7] showing that second-generation offspring of rats eating a high-fat diet during pregnancy had an 80% chance of cancer, compared with 50% of control rats. 'Why you should worry about grandma's eating habits', read one headline. "Think twice about that bag of potato chips because you are eating for more than two," warned another story. These articles did not state that the rats were bred for high cancer rates. Nor did they include inconsistent results: third-generation offspring of female rats on high-fat diets actually had lower incidences of tumours than their control peers.

> **"We urge scientists, educators and reporters to anticipate how this work is likely to be interpreted in popular discussions."**

Inadequately supported and poorly contextualized statements are also found in well-intentioned educational materials. The website beginbeforebirth.org, put together by researchers at Imperial College London, advocates ways to "support and look after pregnant women". A video on the website portrays a 19-year-old released from prison after a stint for looting (see go.nature.com/wynfzw). "Perhaps his problems stretch right back to the womb," the narrator says. "Could better care of pregnant women be a new way of preventing crime?" At best, such suggestions overstate conclusions of current research.

BEYOND THE MATERNAL IMPRINT

Today, an increasing segment of DOHaD research recognizes that fathers and grandparents also affect descendants' health. Studies suggest that diet and stress modify sperm epigenetically and increase an offspring's risk of heart disease, autism and schizophrenia. In humans, the influence of fathers over mothers' psychological and physical state is increasingly recognized. So are effects of racial discrimination, lack of access to nutritious foods and exposure to toxic chemicals in the environment.

Viewed from this broader perspective, DOHaD provides a rationale for policies to improve the quality of life for women and men. It must not be used to lecture individual women, as in a 2014 news report from the US media organization National Public Radio on an epigenetics study in mice: "Pregnancy should be a time to double-down on healthful eating if you want to avoid setting up your unborn child for a lifetime of wrestling with obesity." How are women who lack time or access to healthy foods to act on such advice?

We urge scientists, educators and reporters to anticipate how DOHaD work is likely to be interpreted in popular discussions. Although no one denies that healthy behaviour is important during pregnancy, all those involved should be at pains to explain that findings are too preliminary to provide recommendations for daily living.

Caveats span four areas. First, avoid extrapolating from animal studies to humans without qualification. The short lifespans and large litter sizes favoured for lab studies often make animal models poor proxies for human reproduction. Second, emphasize the role of both paternal and maternal effects. This can counterbalance the tendency to pin poor outcomes on maternal behaviour. Third, convey complexity. Intrauterine exposures can raise or lower disease risk, but so too can a plethora of other intertwined genetic, lifestyle, socio-economic and environmental factors that are poorly understood. Fourth, recognize the role of society. Many of the intrauterine stressors that DOHaD identifies as having adverse intergenerational effects correlate with social gradients of class, race and gender. This points to the need for societal changes rather than individual solutions.

Although remembering past excesses of 'mother-blame' might dampen excitement about epigenetic research in DOHaD, it will help the field to improve health without constraining women's freedom. ■

Sarah S. Richardson *is associate professor of the history of science and of studies of women, gender and sexuality at Harvard University in Cambridge, Massachusetts, USA.* **Cynthia R. Daniels** *is professor of political science at Rutgers University in New Brunswick, New Jersey, USA.* **Matthew W. Gillman** *is professor of population medicine and director of the Obesity Prevention Program at Harvard Medical School in Boston, Massachusetts, USA.* **Janet Golden** *is professor of history at Rutgers University in Camden, New Jersey, USA.* **Rebecca Kukla** *is professor of philosophy at Georgetown University in Washington DC, USA.* **Christopher Kuzawa** *is professor of anthropology at Northwestern University in Evanston, Illinois, USA.* **Janet Rich-Edwards** *is associate professor of medicine at the Connors Center for Women's Health and Gender Biology at Harvard Medical School in Boston, Massachusetts, USA.*
e-mail: srichard@fas.harvard.edu

1. Barker, D., Barker, M., Fleming, T. & Lampl, M. *Nature* **504,** 209–211 (2013).
2. Pettitt, D. J. *et al. Diabetes* **37,** 622–628 (1988).
3. Kaskutas, L. & Greenfield, T. K. *Drug Alcohol Depend.* **31,** 1–14 (1992).
4. Kesmodel, U. S. *et al. BJOG* **119,** 1180–1190 (2012).
5. Frank, D. A., Augustyn, M., Knight, W. G., Pell, T. & Zuckerman, B. *J. Am. Med. Assoc.* **285,** 1613–1625 (2001).
6. Parboosing, R., Bao, Y., Shen, L., Schaefer, C. A. & Brown, A. S. *JAMA Psychiatry* **70,** 677–685 (2013).
7. de Assis, S. *et al. Nature Commun.* **3,** 1053 (2012).

CHRISTIE'S IMAGES/CORBIS

James Eckford Lauder's *James Watt with the Newcomen Engine* (1855), painted after the late engineer had become a celebrated figure.

HISTORY OF ENGINEERING

Wonder maker

Andrew Robinson delves into a study inspired by James Watt's fascinating workshop.

In 1924, London's Science Museum acquired the entire workshop of engineer James Watt, left almost untouched in the attic of his house in Birmingham, UK, since his death more than a century before. The museum put a recreation of the workshop on permanent display in 2011. Among the 8,434 items left by the Scotsman, best known for his innovative steam engine, is an enormous range of tools, including the earliest known circular saws. There are also mathematical instruments, optical experiments, minerals and chemicals, pottery and ceramics made by Watt, busts of famous figures waiting to be copied in plaster of Paris, and engine-related objects — such as a box containing the fragments of his attempts to make an engine that used pure rotary motion.

This workshop inspired Ben Russell, the Science Museum's curator of mechanical engineering, to write his engaging *James Watt: Making the World Anew*. He explains that the volume of material, "crossing the boundaries between philosophy and craft, makes it hard to categorize the contents against any one of the labels which have been applied to Watt over time: philosopher or craftsman primarily, but engineer and chemist, as well." The diversity of Watt's interests and activities was astonishing, even when compared with the achievements of his Enlightenment contemporaries. Chemist, inventor and Royal Society president Humphry Davy, for instance, called him a "modern Archimedes" whose inventions had made industrialized Britain remarkably powerful for such a small nation.

James Watt: Making the World Anew
BEN RUSSELL
Reaktion: 2014.

Watt's first steam engine, which began operating in 1776, was successful because it had three times the coal-combustion efficiency of the existing engine designed by Thomas Newcomen, introduced in 1712. The steam cylinder in Newcomen's 'atmospheric' engine had to be sprayed with cold water at each cycle to condense the steam, creating a partial vacuum that allowed atmospheric pressure to push the piston down. In 1765, in Glasgow, Watt had a "major leap of imagination", as Russell puts it: the idea of building a separate condenser, so that cylinder and piston did not lose heat. By patenting the principles of the condenser and not the means of applying them, Watt and his business partner Matthew Boulton became wealthy, although not without a long legal battle against their rivals in the 1790s. Their engine — its power defined in horsepower, a unit invented by Watt and today most commonly converted as 746 watts — became an industry standard by 1800, for pumping water from mines and driving machinery in mills and factories.

From 1804, Watt moved from steam to sculpture, creating plaster of Paris copies of busts, then much in demand among the wealthy. His 'sculpture machine' was a three-dimensional pantograph, powered by a treadle and worked by means of linked and geared arms, one ending in a probe and one in a high-speed, rotating cutting tool. As the probe traced the surface of the original bust, the tool duplicated its motion and cut a plaster block. Today, about 400 of Watt's sculptures are in storage at the Science Museum, including casts, busts, depictions of contemporaries including the chemist Joseph Black, and copies of Boulton's 1809 death mask. After his own death in 1819, Watt became the first engineer to be commemorated in Westminster Abbey. For the Victorians, Russell shows, Watt was "a new kind of industrial hero" whose stature was comparable to Isaac Newton's as a physicist.

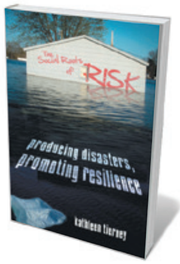
As Russell admits, there is no shortage of recently published studies of Watt, such as Richard Hills's three-volume biography *James Watt* (Landmark, 2002–06) and *James Watt, Chemist* by David Miller (Pickering & Chatto, 2009). But where Russell focuses on Watt as a man able "not just to think but to do: to use tools, techniques and materials, to create tangible things across a range of activities", most studies tend to emphasize his capacity as a thinker. Perhaps that tendency is inevitable. Scientists and science historians generally revere original theories with unforeseeable consequences more than practical inventions with immediate applications — Newton and Albert Einstein more than Christopher Wren, Watt and Thomas Edison. For all the wonderful creativity on display in his workshop, Watt was essentially earthbound. Yet his life and work are decidedly relevant to the debate about how scientific discoveries are best turned into marketable inventions. Watt's way of working — with a business partner and a patentable purpose, whether an efficient coal-driven means of pumping flood water out of mineshafts or the mass production of pottery — could hold lessons for any university or government keen to promote technology transfer.

"Watt was 'a new kind of industrial hero' whose stature was comparable to Isaac Newton's as a physicist."

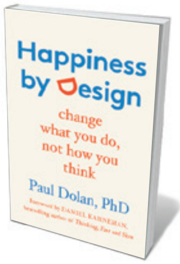
Watt was born in Scotland, trained as an instrument maker in England, made his breakthrough with the steam engine in Scotland, and began manufacturing it in England, where he settled. Next month, there will be a referendum on Scottish independence from the United Kingdom. Whatever the outcome, Watt's remarkable life is a definite benefit arising from the close economic, intellectual and cultural union of Scotland and England. ■

Andrew Robinson *is the author of* The Last Man Who Knew Everything *— a biography of the polymath Thomas Young — and editor of* The Scientists.
e-mail: andrew.robinson33@virgin.net

Books in brief

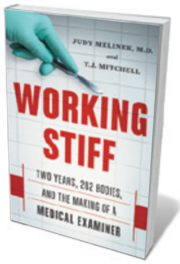


The Social Roots of Risk: Producing Disasters, Promoting Resilience
Kathleen Tierney Stanford University Press *(2014)*
The origins of disaster lie in "the ordinary everyday workings of society", avers sociologist Kathleen Tierney in this brilliant treatise. Drawing on a trove of timely case studies, Tierney analyses how factors such as speculative finance and rampant development allow natural and economic blips to tip more easily into catastrophe. Resilience, she argues, is rooted in sustainable ecological and social development. It is transformative risk reduction, not bailouts, that will help humanity to weather coming upheavals.

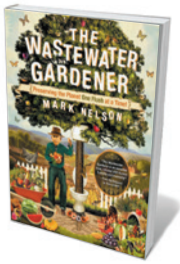


Happiness by Design: Change What You Do, Not How You Think
Paul Dolan Hudson Street *(2014)*
The science of happiness has been with us since at least the 1940s, when Abraham Maslow's ideas opened up a psychology based on feeding the potential for positivity rather than simply treating symptoms. To this now-crowded table, behavioural scientist Paul Dolan brings a feast of US and European research, and some significant insights. Dolan argues that happiness depends on where we focus our attention, and on how well we balance purpose and pleasure. His action-oriented outline for achieving that equilibrium draws in part on work with eminent psychologist Daniel Kahneman.

Great Minds: Reflections of 111 Top Scientists
Balazs Hargittai, Magdolna Hargittai and Istvan Hargittai
Oxford University Press *(2014)*
Over two decades, chemists Balazs, Magdolna and Istvan Hargittai interviewed hundreds of prominent scientists, including 68 Nobel laureates. This distillation features excerpts from 111 of these frank conversations. Featured are mathematician John Conway on how his discovery of surreal numbers was like finding a palace after drifting around a strange city; physicist Gerard 't Hooft on the improbability of intelligent extraterrestrials; and physicist Mildred Dresselhaus, biologist Francis Crick, and more on the fascination of the life scientific.



Working Stiff: Two Years, 262 Bodies, and the Making of a Medical Examiner
Judy Melinek and T. J. Mitchell Scribner *(2014)*
"A hard hat was still there, lying on its side in a pool of blood and brains, coffee and doughnuts." Judy Melinek's inside story on forensic-pathology training, written with her husband, writer T. J. Mitchell, is inevitably big on gore. But Melinek, a "sunny optimist", offers more than cheap thrills. The flamboyant disclosures — how to handle rotting flesh or use pruning shears to snap ribs — are balanced by her soul-baring account of identifying human remains in the wake of the terrorist attacks in New York on 11 September 2001.



The Wastewater Gardener: Preserving the Planet One Flush at a Time
Mark Nelson Synergetic *(2014)*
It takes 1,000 tonnes of water to move 1 tonne of human faeces, notes engineer Mark Nelson. His alternative to costly, unsustainable sanitation is constructed wetland — subsurface-flow gravel beds in which plant roots and microbial action purify wastewater for a full range of uses. Nelson, a veteran of the 1990s US survivability experiment Biosphere 2, has built "wastewater gardens" from Algeria to Australia, Mexico and beyond. **Barbara Kiser**

Correspondence

Russian stamp to honour physicist

Russia has just issued a postage stamp to mark the centenary of the birth of the brilliant physicist and cosmologist Yakov Zel'dovich (1914–87).

Among his many achievements, and despite never having received a university degree, Zel'dovich developed the theories of nuclear chain reactions and of the gravitational lens (see R. A. Sunyaev (ed.) *Zel'dovich: Reminiscences* Chapman & Hall/CRC; 2004).

As a theoretician, he was involved in the creation of Soviet nuclear weapons — the atomic bomb in 1949, with Lev Landau, and the hydrogen bomb in 1953, with Andrei Sakharov. Like Robert Oppenheimer in the United States, he met with government opposition when he declined to continue working on weapons development.

Moving over to astrophysics, Zel'dovich made seminal contributions in gravitational instabilities and in cosmological fluctuations, with the Sunyaev–Zel'dovich effect being among the best known. In 2001, an asteroid, 11438 Zel'dovich, was named in his honour.

Renad I. Zhdanov *Kazan Federal University; and Sholokhov Moscow State University for the Humanities, Russia.*
Pascal Chardonnet *University of Savoie, Annecy, France.*
zrenad@gmail.com

Я.Б. ЗЕЛЬДОВИЧ
1914–1987
РОССИЯ RUSSIA·2014
15 Р.

MARKA PUBLISHING AND TRADING CENTRE

White possums must stay cool to survive

It is ironic that Australia, one of the world's highest carbon emitters per capita, is giving up on a hard-won plan to reduce its greenhouse-gas emissions (see *Nature* **511,** 392; 2014) just as climate change could be about to claim one of its rarest and most iconic animals — the white lemuroid ringtail possum (*Hemibelideus lemuroides*).

The white possum was once abundant in cool rainforests on Mount Lewis in northern Queensland, but its population collapsed abruptly following a severe heatwave in 2005. Today, just a handful of individuals are left (see J. Chandler *New Scientist* Issue 2980, 42–45; 2014).

Tropical mountains are full of endemic species that have adapted to cooler local climates and are particularly vulnerable to heatwaves and other extreme weather associated with climate change (see go.nature.com/vqskfa).

It has therefore been suggested that the white possum might be a more sensitive indicator of climate change than the polar bear (W. Laurance *New Scientist* Issue 2690, 14; 2009). But first the temperature of its habitat must be stabilized so that its numbers can be restored.

William F. Laurance, Susan Laurance *James Cook University, Cairns, Australia.*
bill.laurance@jcu.edu.au
Christine Milne *Parliament House, Canberra, Australia.*

Mexican GM maize rift is not so simple

You rightly point out that the issue of genetically modified (GM) maize (corn) is more sensitive and complex in Mexico than in other countries (*Nature* **511,** 16–17; 2014), but you owe readers a more in-depth and balanced view.

The rift in Mexico's scientific community over GM maize is not directly related to the legal challenge you discuss. It is a result of the commercial push to plant GM maize before the benefits and risks, and the costs to Mexican society, have been fully assessed.

The possibility of producing maize that is tolerant to drought and frost, a claim you report from government-funded researchers, could indeed help to restore Mexico's capacity for growing its own maize. However, commercial cultivars in Mexico (25% of total area) have limited reach, even after more than 60 years of breeding (see, for example, S. Brush and H. Perales *Agr. Ecosyst. Environ.* **121,** 211–221; 2007). More than two million households rely on traditional landraces for food security (H. Eakin *et al. Dev. Change* **45,** 133–155; 2014), and the global prevalence of insecticide-producing and herbicide-tolerant GM products is at more than 98% after almost 20 years (see go.nature.com/jyux8p). These factors mean that such claims need to be realized and qualified if they are to be taken seriously.

Those seeking commercial acceptance of GM maize still need to convince key groups in Mexican society, including scientists, that the benefits of planting it will outweigh the risks and social costs. There is more to maize in Mexico than productivity and business, and it is not only scientists and seed companies who have rights.

Hugo Perales *El Colegio de la Frontera Sur (ECOSUR), San Cristóbal, Chiapas, Mexico.*
hperales@ecosur.mx

Create ethics codes to curb sex abuse

A survey published last month found evidence of alarming levels of sexual violence (towards 26% of women and 6% of men) in the course of fieldwork by life scientists (see *Nature* http://doi.org/t3n; 2014). Meanwhile, more than 50 US higher-education institutions are under investigation for their handling of complaints of such incidents. As a rape survivor and scientist, I suggest measures that could help to counteract this situation.

Scientific research organizations should draw up professional codes of ethics, akin to those of the Modern Language Association of America and the American Historical Association, with explicit provisions that denounce sexual harassment and discrimination on the basis of race, gender or sexual orientation.

A national framework that academic, industrial and government institutions could sign or adapt would be an important step. Such proactive strategies would prevent interference with the core work of researchers. In the United Kingdom, for example, the Athena SWAN Charter outlines a series of best practices to further and protect women's careers (see go.nature.com/mkxlr8).

Institutions must make clear the repercussions for students and employees who transgress, and provide a mechanism for consistent enforcement (such as an adjudication committee) that would complement any legal redress.

Margaret C. Hardy *University of Queensland, Brisbane, Australia.*
m.hardy@imb.uq.edu.au

NEUROSCIENCE

What females really want

The identification of neural subcircuits used by female fruit flies to make a choice about whether to copulate with a potential mate provides a template for understanding how the brain integrates complex information to reach decisions.

LESLIE C. GRIFFITH

For humans, the decision to copulate is an intensely personal one, and we like to believe it is a choice made under free will. However, more than a century of studying other species has made it clear that specific internal states, together with the presence of particular external cues, can alter the probability of copulation in a consistent way across a population, strongly suggesting that there are neural circuits that evaluate relevant, predetermined variables and so bias behaviour. Decision-making circuits are present in all species with a nervous system, and understanding how the brain carries out this type of computation is a major goal of neuroscience. Now, three studies (one published in the journal *Current Biology*[1] and two in *Neuron*[2,3]) using different genetic strategies have identified circuit components that control the receptivity of female fruit flies to male courtship, outlining the scope of this complex decision-making process.

Sexual behaviour in the fruit fly *Drosophila melanogaster* is a particularly useful model system for studying decision-making because it involves both stereotyped and plastic features. The complicated sex-related behaviour of flies is conducted by a relatively small brain, and investigators have an arsenal of sophisticated genetic tools with which to reproducibly identify and manipulate particular neurons in freely behaving animals[4]. Although the courtship behaviour of males has been studied intensely, progress in understanding female reproductive behaviour[5,6] has made it apparent that females are not simply passive recipients of male advances. Instead, female flies engage in an active and complex decision-making process[5,6] that determines whether copulation occurs. The female's decision-making apparatus uses sensory information — including courtship songs produced by male wing vibration, visual cues and olfactory cues such as pheromones — to assess male fitness in the context of the internal state of the female herself (Fig. 1).

Bussell *et al.*[1] undertook a genome-wide screen to look for genes that alter the receptivity of female flies to potential mates. They found that decreasing the activity of the transcription factor Abdominal-B (Abd-B) in female neurons decreased the rate of mating. The authors showed that neurons expressing Abd-B during development regulate the rate of female pausing during courtship, an indicator of receptivity. Abd-B-expressing neurons were activated in response to male-specific sensory inputs such as courtship song, but only in the presence of male flies (playback of a recording of the song alone was ineffective, indicating that the male probably provides additional visual or chemosensory cues). The fact that these neurons are downstream of sensory inputs suggests that Abd-B neurons are part of the receptivity-output arm of the fly decision-making circuitry, and are driven by higher centres that process and integrate courtship-relevant information.

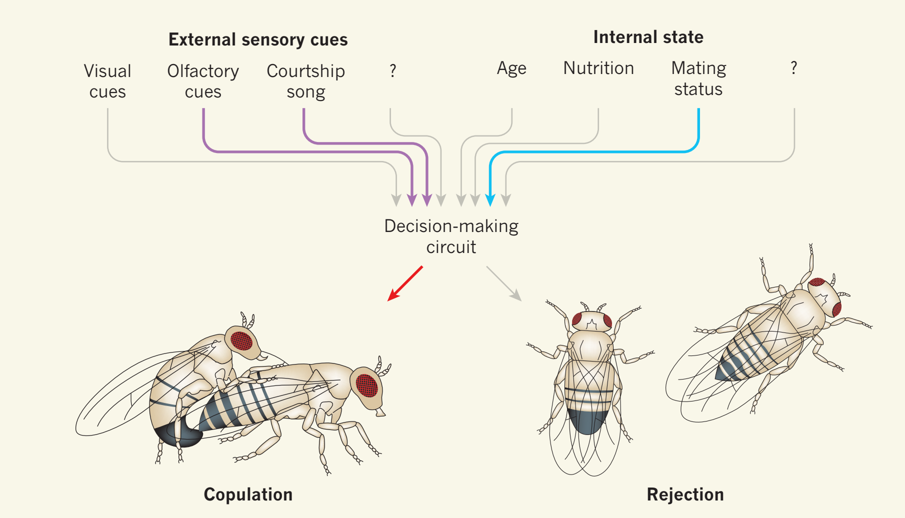


Figure 1 | To mate or not to mate. When making a decision about whether to copulate with a potential mate, a female fruit fly processes external sensory cues provided by the candidate male and the environment during the courtship ritual, in addition to information about her own internal state. These various inputs are integrated into decision-making circuits, and the decision is relayed by regulatory output circuits. Three studies have identified neurons involved in this process: Zhou *et al.*[2] identified the input neurons for olfactory cues and courtship songs (purple arrows); Feng *et al.*[3] identified the input neurons for mating status (blue arrow); and Bussell *et al.*[1] identified the output circuit leading to copulation (red arrow). The neurons related to various other inputs, such as visual cues, remain to be identified (grey arrows).

In their hunt for circuits involved in this sensory integration, Zhou *et al.*[2] began with the assumption that neurons that express *doublesex* (*dsx*), a gene that is differentially expressed in male and female reproductive circuits, would contribute to female-specific behaviour[6]. The authors used state-of-the-art genetic techniques[4] to identify and manipulate small populations of neurons expressing Dsx protein in the female adult brain, and found that activation of two neuron groups, pC1 and pCd, enhanced the rate of copulation.

Zhou and colleagues provide anatomical evidence to suggest that pC1 and pCd convey information between brain areas known to be involved in processing courtship-related sensory information[5,6]. Using calcium levels as an indicator of neuronal activation, the authors showed that pCd was responsive to *cis*-vaccenyl acetate, a male-specific lipid pheromone that enhances female receptivity. Male song activated pC1, and this response was enhanced by the presence of *cis*-vaccenyl acetate, suggesting that pC1 neurons convey integrated information. Thus, pC1 and pCd are part of the central circuitry that processes courtship-related information.

In addition to extrinsic cues, receptivity is dependent on the female's internal state. Females that have recently mated do not copulate, even if presented with a fit and eager male.

This change in behaviour is due to transfer of the protein sex peptide (SP) to the female in the male's seminal fluid. SP-responsive receptors have been identified[7] on nerve cells in the abdominal ganglion (a neuronal structure roughly analogous to the spinal cord) that innervate the reproductive tract, but it was unknown how the SP signal is transmitted to the central nervous system.

Feng *et al.*[3] looked for groups of neurons that decreased female receptivity when electrically silenced. Their screen identified a group of neurons that project from the abdominal ganglion into the brain, which they named SP abdominal ganglion (SAG) neurons. The authors observed that these neurons are not themselves responsive to SP, but rather receive information from SP-sensitive neurons through synapses (junctions that transfer signals between cells). Crucially, the strength of this connection was modulated by SP and correlated with the female's mating status. These data establish SAG neurons as a conduit of information on mating status from the reproductive tract to the central nervous system.

Except to those of us who are dedicated fly voyeurs, the importance of these individual studies might be debatable. In aggregate, however, they provide a framework that could lead to a detailed cellular and molecular understanding of a multifactorial decision-making process. By highlighting three different and as-yet-unconnected regions of the female fly's sexual-behaviour circuitry, these studies provide starting points for completing the wiring diagram.

Flies are faced with many of the same basic challenges as humans: what to eat, when to sleep and whom to mate with. Choice — our exercise of free will — is the probabilistic representation of integration processes that are rooted in the molecular and neural architecture of our brains. The genetic and electrophysiological tools available in the fly make this model organism arguably the best place to get our first glimpse into how a brain can make complex decisions. ■

Leslie C. Griffith *is in the Department of Biology, Volen Center for Complex Systems and National Center for Behavioural Genomics, Brandeis University, Waltham, Massachusetts 02454-9110, USA.*
e-mail: griffith@brandeis.edu

1. Bussell, J. J., Yapici, N., Zhang, S. X., Dickson, B. J. & Vosshall, L. B. *Curr. Biol.* **24,** 1584–1595 (2014).
2. Zhou, C., Pan, Y., Robinett, C. C., Meissner, G. W. & Baker, B. S. *Neuron* **83,** 149–163 (2014).
3. Feng, K. Palfreyman, M. T., Häsemeyer, M., Talsma, A. & Dickson, B. J. *Neuron* **83,** 135–148 (2014).
4. Venken, K. J. T., Simpson, J. H. & Bellen, H. J. *Neuron* **72,** 202–230 (2011).
5. Ferveur, J.-F. *Curr. Opin. Neurobiol.* **20,** 764–769 (2010).
6. Pavlou, H. J. & Goodwin, S. F. *Curr. Opin. Neurobiol.* **23,** 76–83 (2013).
7. Yapici, N., Kim, Y.-J., Ribeiro, C. & Dickson, B. J. *Nature* **451,** 33–37 (2008).

SOLAR SYSTEM

Sandcastles in space

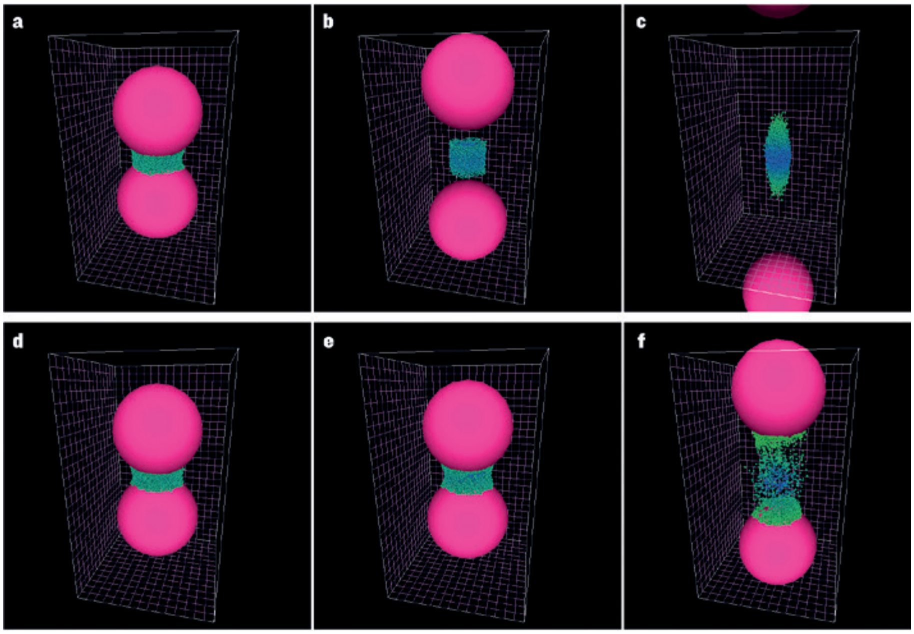
Analysis of a kilometre-sized, near-Earth asteroid shows that forces weaker than the weight of a penny can keep it from falling apart. This has implications for understanding the evolution of the Solar System. SEE LETTER P.174

DANIEL J. SCHEERES

Our logical concepts for how asteroids should behave have taken another knock, as evidenced in a paper by Rozitis *et al.*[1] on page 174 of this issue. The researchers establish that a kilometre-sized, near-Earth asteroid known as (29075) 1950 DA is covered with sandy regolith (the surface covering of an asteroid) and spins so fast — one revolution every 2.12 hours — that gravity alone cannot hold this material to its surface. This places the asteroid in a surreal state in which an astronaut could easily scoop up a sample from its surface, yet would have to hold on to the asteroid to avoid being flung off.

Rozitis and colleagues show that for this rubble-pile asteroid (the body has a porosity of roughly 51%) to stay in one piece, it must have cohesive strength — just not very much. On the basis of the density, size and shape of 1950 DA, the authors find that the asteroid requires a cohesive strength of at least 64 pascals to hold all of its rubble-pile components together: similar to the pressure that a penny exerts on the palm of your hand.

This strength is consistent with, but much more precisely determined than, similar levels of cohesive strength that have been deduced for rubble-pile asteroids on the basis of spin-rate and size statistics of asteroids[2] and on the inferred strength, size and spin rate of the active asteroid P/2013 R3 (ref. 3). This asteroid was recently observed to comprise several chunks that are slowly escaping from each other, probably owing to rotational disruption[4]. A model for how to generate such a modest level of strength in geophysical bodies has been hypothesized[2], and achieves this



REF. 2/THE METEORITICAL SOCIETY (2014)

Figure 1 | Cohesive forces in regolith. Computer simulations[2] of two metre-sized boulders (pink spheres) with loosely packed centimetre-sized regolith (green and blue particles) between them. The whole system is under self-gravitational attraction, and to determine its strength, the boulders are pulled apart with an increasing force. Panels **a** and **d** show the initial configuration, with the pulling force exactly equal to the gravitational attraction. Panels **b, e,** and **c, f** show the system response for equal forces beyond the gravitational limit. If the regolith has no cohesive strength (panels **a**–**c**), it immediately separates from the boulders once they are pulled with a force greater than their gravitational attraction, which leaves the regolith behind to aggregate under its own self-gravity and provides no extra strength to the system. If there are cohesive van der Waals forces between the regolith particles (panels **d**–**f**), the particles serve as a glue and strengthen the bond between the boulders. The level of cohesion required to hold rubble-pile asteroid (29075) 1950 DA together, found by Rozitis *et al.*[1] to be 64 pascals, can be generated by a loosely packed regolith with particles as fine as roughly 10 micrometres[2].

through 'dry cohesion' arising from van der Waals forces between components of a rubble pile (Fig. 1). In this theory, the finest grains (potentially as small as 1–10 micrometres) in a rubble pile that are present in sufficient numbers to connect all larger grains provide a very weak cement that can hold the body together — a fairy-dust version of a sandcastle.

Although this image of fairy-castle asteroids is entertaining, the implications of these measurements are far-reaching. A defining feature of the rubble pile 1950 DA is that it is globally in a microgravity environment — the centrifugal forces from its rapid spin rate are nearly balanced by its gravitational attraction, with the difference between them being a tiny fraction of Earth's gravity. In such a regime, weak van der Waals forces can dominate[5]. The evident stability of such a strange body as 1950 DA exposes our ignorance of how the geophysics of asteroids works in the microgravity regime, with its current state being difficult to reconcile with classical views of how rubble-pile bodies form from catastrophically disrupted parent bodies. Although Rozitis *et al*. lay out a plausible story for the current state of 1950 DA, the development of a complete theory of microgravity geophysics could have significant consequences, beyond this single case, for our evolving understanding of asteroids and the Solar System.

For asteroids, the larger implications of such a weakly cohesive material — for example, the dissipation of energy in their interiors[6], the shedding of material from their surfaces[7] and the creation of binary asteroid systems through the fissioning of rapidly rotating rubble piles[8,9] — have yet to be fully explored and understood. Going beyond asteroids, many different bodies and environments in the past and present Solar System lie in microgravity regimes similar to that of 1950 DA, where inertial, gravitational and weak molecular forces may be simultaneously relevant. The effects of the interplay of these forces in, for instance, the creation and destruction of transient structures in planetary ring systems and the accretion of grains in protoplanetary disks all become ripe topics for investigation motivated by this example.

Coming back to near-Earth asteroids, this result and the underlying theory also have ramifications for the exploration of small asteroids such as 1950 DA, currently a topic of great interest to national space agencies and a few private corporations. Small amounts of cohesion in an asteroid's regolith can enable its surface to become 'perched', just waiting for a meteorite impact (or passing astronaut) to destabilize it — similar to avalanches on Earth. The global strength of such rubble-pile asteroids held together by these weak forces is also unclear. How often might avalanches consume the entire body, causing it to split and disassemble? Recent observations of active asteroids seem to indicate that such natural outcomes might not be that rare[4,7].

The ability for human or robotic interactions to create such global changes to a small asteroid suggests an intriguing vision of geophysical laboratories in space. Given that small, near-Earth asteroids are accessible using spacecraft, it becomes possible to do controlled geophysical experiments on these bodies that result in global and locally measurable changes. This would allow us to probe the geophysics of microgravity aggregates in their natural environments, and to do so at scales that cannot be recreated on Earth or in Earth's orbit, at the cost of a modest planetary-science mission.

Independent of whether we choose to take advantage of such natural laboratories in the near future, humanity might eventually have no choice, because 1950 DA is due to pay an uncomfortably close visit to Earth. The asteroid is one of the most potentially hazardous known, with a 1 in 4,000 chance of impacting the Earth in the year 2880 (ref. 10). Such an impact could have planet-wide consequences owing to the asteroid's size. Among the many proposed methods for deflecting this hazard is to run a massive spacecraft into it at high speed, or to set off a nuclear blast in close proximity[11]. However, for this weakly bound body, we should wonder whether such an attempt would make it crumble and fall apart like a sandcastle that has been baked in the sunshine.

Whether the impact from such a disaggregated asteroid would pose a larger threat to Earth has been a matter of debate in the scientific community. Whereas a single asteroid packs a larger punch, the shotgun spray from a disaggregated body may hit multiple sites across the globe. For a rapid rotator such as 1950 DA, however, this is not a relevant question. Once released from each other, the speed of the body's components relative to the asteroid's centre of mass would range from tens of centimetres per second if it split in half, to up to 50 cm s^{-1} for material that might break off its surface. These speeds are much greater than most mitigation techniques could deliver to the parent asteroid. This would cause the components to drift relative to the initial impact trajectory by more than one Earth radius in less than a year — sparing humanity from having to resolve such a delicate question. ■

Daniel J. Scheeres *is in the Department of Aerospace Engineering Sciences, The University of Colorado, Boulder, Colorado 80309, USA.*
e-mail: scheeres@colorado.edu

1. Rozitis, B., MacLennan, E. & Emery, J. P. *Nature* **512,** 174–176 (2014).
2. Sánchez, P. & Scheeres, D. J. *Meteorit. Planet. Sci.* **49,** 788–811 (2014).
3. Hirabayashi, M., Scheeres, D. J., Sánchez, D. P. & Gabriel, T. *Astrophys. J. Lett.* **789,** L12 (2014).
4. Jewitt, D. *et al. Astrophys. J. Lett.* **784,** L8 (2014).
5. Scheeres, D. J., Hartzell, C. M., Sánchez, P. & Swift, M. *Icarus* **210,** 968–984 (2010).
6. Goldreich, P. & Sari, R. *Astrophys. J.* **691,** 54 (2009).
7. Jewitt, D., Agarwal, J., Weaver, H., Mutchler, M. & Larson, S. *Astrophys. J. Lett.* **778,** L21 (2013).
8. Walsh, K. J., Richardson, D. C. & Michel, P. *Nature* **454,** 188–191 (2008).
9. Jacobson, S. A. & Scheeres, D. J. *Icarus* **214,** 161–178 (2011).
10. Farnocchia, D. & Chesley, S. R. *Icarus* **229,** 321–327 (2014).
11. Ahrens, T. J. & Harris, A. W. *Nature* **360,** 429–433 (1992).

AGEING

Old blood stem cells feel the stress

Ageing is accompanied by deterioration in the haematopoietic stem cells that are responsible for regenerating the blood system. Cellular stress in the aged stem cells could be a cause of this decline. SEE LETTER P.198

JIRI BARTEK & ZDENEK HODNY

Tissue renewal is a fundamental process that relies on the regenerative capacity of long-lived, self-renewing stem cells. But during ageing, stem-cell function deteriorates. The haematopoietic stem cells (HSCs) that maintain all blood-cell lineages are, like other long-lived stem cells, prone to accumulating DNA damage as they age. In the case of HSCs, the damage can reduce the cells' ability to regenerate blood-cell lineages, and can increase the risk of diseases such as leukaemia. But little is known about what causes the damage and how it contributes to the decline of old HSCs. On page 198 of this issue, Flach *et al.*[1] report that damage is caused mostly by cellular stress that arises as a result of inefficient DNA replication, and they point to the probable molecular defects involved.

DNA damage occurs when cells cannot repair genetic inaccuracies, which frequently

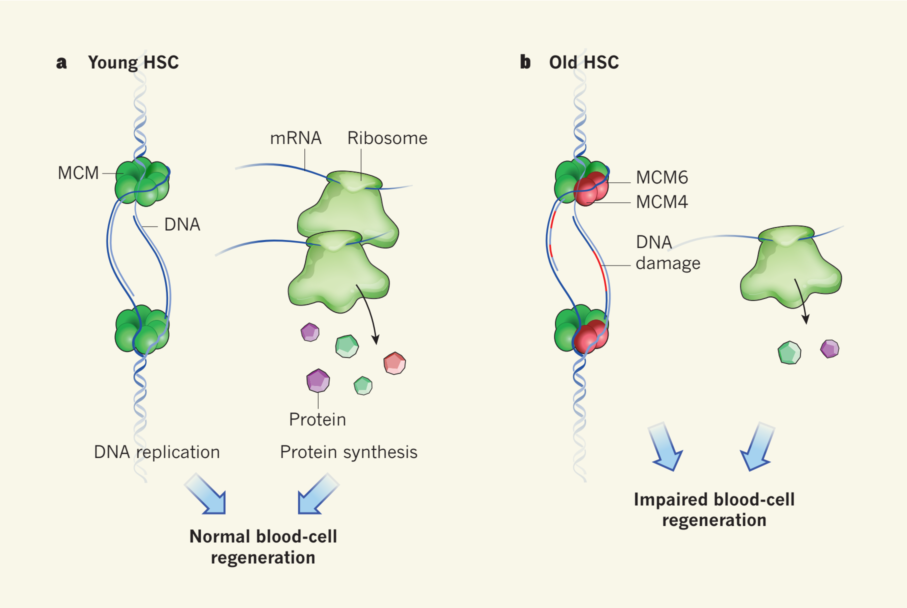


Figure 1 | Replication stress in ageing cells. **a**, In young haematopoietic stem cells (HSCs), the MCM protein complex promotes DNA replication, whereby new DNA strands are generated. In a separate process, the messenger RNA (mRNA) produced during gene transcription moves to a structure called the ribosome to be translated into proteins. These processes enable HSCs to self-renew and give rise to all blood-cell lineages. **b**, Flach *et al.*[1] report that aged HSCs have lower levels of two subunits of MCM, MCM4 and MCM6, than do young cells, which prevents the MCM complex from working properly, resulting in DNA replication stress. DNA damage associated with replication stress is not properly repaired in old HSCs, which leads to abnormalities in genes encoding ribosomal components, impaired ribosome assembly and reduced protein production. These combined stresses on HSCs lead to abnormal production of blood-cell lineages.

arise while DNA is being replicated during cell proliferation. The idea that DNA damage is a major driver of the deterioration of stem cells in general, and old HSCs in particular, is supported by the fact that both mice and people with deficiencies in DNA repair age more quickly than those without such deficiencies[2–5]. But debate over the potential causes of DNA damage in HSCs has been lively and multifaceted, because factors both intrinsic to the cell itself (for example, loss of cell polarity) and extrinsic (such as secreted proteins or changes in the types of cell surrounding the HSCs) can affect the environment in which old HSCs reside[6].

To investigate the origin and impact of DNA damage in aged HSCs, Flach and colleagues compared HSCs isolated from the bone marrow of young and old mice. Compared with young cells, old HSCs showed a functional decline, together with signalling indicative of DNA damage, which the authors gauged by presence of the γH2AX protein. γH2AX was accompanied by an increased abundance of proteins associated with inefficient DNA replication (known as DNA replication stress)[7]. These proteins promote signalling by the enzyme ATR, which modifies many cellular functions[3,7].

Following up on this unexpected result, the authors found that ATR signalling was activated in old HSCs, another indication that they were subject to replication stress. The cells also showed delayed entry into and progression through S phase, the period of the cell cycle in which the genome is replicated. Furthermore, DNA replication frequently stalled in old HSCs, and the number of 53BP1 bodies — structures that mark chromosomal breaks in the nuclei of cells that have experienced replication stress[8] — rose.

To look at what molecular defects could be responsible for enhanced replication stress in aged HSCs, Flach *et al.* compared gene-expression profiles in young and old HSCs. Genes encoding the proteins MCM4 and MCM6 (two components of an MCM protein complex that is essential for proper replication) showed lower expression in old than young HSCs, as did a variety of other factors.

The authors found that experimental depletion of MCM4 and MCM6 in young HSCs impaired the cells' function. Like old HSCs, the altered cells had a poor capacity to regenerate the blood system when transplanted into mice, suggesting that low levels of MCM4 and MCM6 are linked with replication stress, and thereby with functional deterioration. In agreement with this, young HSCs were also impaired if replication stress was caused by chemical compounds.

Finally, Flach *et al.* investigated why γH2AX was present in HSCs that had stopped proliferating and therefore could not be experiencing replication stress. They found signs of long-term damage signals in genes within ribosomal DNA (rDNA), which includes many genes that encode components involved in assembly of the ribosome (the cellular machinery responsible for producing protein from messenger RNA). This makes sense, because rDNA is difficult to replicate and is therefore prone to replication stress. The authors showed that persistent damage was linked to lowered expression of rDNA genes. Consequently, the cells made fewer ribosomes, and could not produce enough protein to sustain cellular function — a state known as ribosomal biogenesis stress[9].

Overall, Flach and colleagues' work shows that old HSCs experience both replication stress and ribosome biogenesis stress. The former probably triggers the latter, and is clearly at least partly responsible for impaired blood regeneration in advanced age (Fig. 1). The results have broad implications for medicine, and raise many questions. For example, is replication stress involved in the deterioration of ageing stem cells in other tissues? Is the authors' mechanism relevant to human HSCs?

Because replication stress underlies many tumours[10], it is possible that stress in HSCs contributes to the progressive accrual of gene mutations that cause ageing-related cancers of the blood. It will be interesting to determine how ribosome biogenesis stress influences HSC decline, and to investigate whether the p53 tumour-suppressor protein — a known sensor of both replication and ribosomal stress[3,9,10] — is involved.

Finally, could restoration of MCM4 and MCM6 levels avert replication stress or even functional decline in old HSCs? If it could, understanding how MCM genes are inhibited in old age might be a good starting point for defining strategies to postpone, prevent or even reverse the deterioration of the ageing blood-regeneration system. ■

Jiri Bartek *and* **Zdenek Hodny** *are at the Institute of Molecular Genetics of the ASCR, v.v.i., Prague 142 20, Czech Republic.* **J.B.** *is also at the Danish Cancer Society Research Center, Genome Integrity Unit, Copenhagen DK-2100, Denmark.*
e-mails: jb@cancer.dk; hodny@img.cas.cz

1. Flach, J. *et al. Nature* **512,** 198–202 (2014).
2. Behrens, A., van Deursen, J. M., Rudolph, K. L. & Schumacher, B. *Nature Cell Biol.* **16,** 201–207 (2014).
3. Jackson, S. & Bartek, J. *Nature* **461,** 1071–1078 (2009).
4. Rossi, D. J. *et al. Nature* **447,** 725–729 (2007).
5. Nijnik, A. *et al. Nature* **447,** 686–690 (2007).
6. Geiger, H., Denkinger, M. & Schimbeck, R. *Curr. Opin. Immunol.* **29,** 86–92 (2014).
7. Zeman, M. K. & Cimprich, K. A. *Nature Cell Biol.* **16,** 2–9 (2013).
8. Lukas, C. *et al. Nature Cell Biol.* **13,** 243–253 (2011).
9. Golomb, L., Volarevic, S. & Oren, M. *FEBS Lett.* **588,** 2571–2579 (2014).
10. Halazonetis, T. D., Gorgoulis, V. & Bartek, J. *Science* **319,** 1352–1355 (2008).

This article was published online on 30 July 2014.

CONDENSED-MATTER PHYSICS

Glasses made from pure metals

The experimental realization of amorphous pure metals sets the stage for studies of the fundamental processes of glass formation, and suggests that amorphous structures are the most ubiquitous forms of condensed matter. SEE LETTER P.177

JAN SCHROERS

On page 177 of this issue, Mao and colleagues[1] report a method that allows them to achieve a long-standing goal for materials scientists — the formation of glasses from pure metals. This will enable much-needed studies of glass formation in simple systems, and allows computational modelling of the processes involved.

For thermodynamic reasons, most liquids become crystalline when they are cooled below their 'liquidus' temperature, above which substances are completely liquid. Crystallizations occur at different timescales and can be suppressed by fast cooling of a liquid, causing it to vitrify into a glass[2]. Vitrification occurs for various materials at widely different critical cooling rates (R_c) — the minimum rate of cooling required to form a glass.

Glass formation has been reported for metallic alloys[3]. An alloy's glass-forming ability increases with the number of components in the alloy, particularly if it contains elements with atomic sizes that differ by more than 12% and which have the thermodynamic impetus to mix[4]. Some alloys that exhibit these criteria, known as bulk metallic glasses, have remarkably good glass-forming abilities, with R_c values of less than 1,000 kelvin per second (comparable with the cooling needed to make amorphous polymers). They also have critical casting thicknesses — the largest thickness over which heat can be extracted enough to avoid crystallization — exceeding 1 millimetre. So far, hundreds of complex alloys have been reported to form bulk metallic glasses.

Pure metals do not fulfil the above criteria because they lack the complexity needed to 'confuse' crystallization[5]. They have therefore been considered to be poor glass formers[6]. Even advanced rapid-cooling techniques have been too slow to avoid crystallization of liquid pure metals, except in some specific cases[7]. Mao and co-workers now introduce a general ultra-rapid heating and cooling method that allows liquids of pure metals to be vitrified.

The authors used a nanometre-scale heating device that brings together two metal tips approximately 100 nm in length. Heating was accomplished using a short electrical pulse (about 4 nanoseconds in duration), which rapidly melted the tips. The heat then dissipated rapidly through the melted sample towards the device, inducing cooling rates of roughly 10^{14} kelvin per second at the centre of the sample. Such high cooling rates were predicted by the researchers to occur on the basis of molecular-dynamics modelling, and caused vitrification of a region of pure metals approximately 40 nm by 50 nm in size.

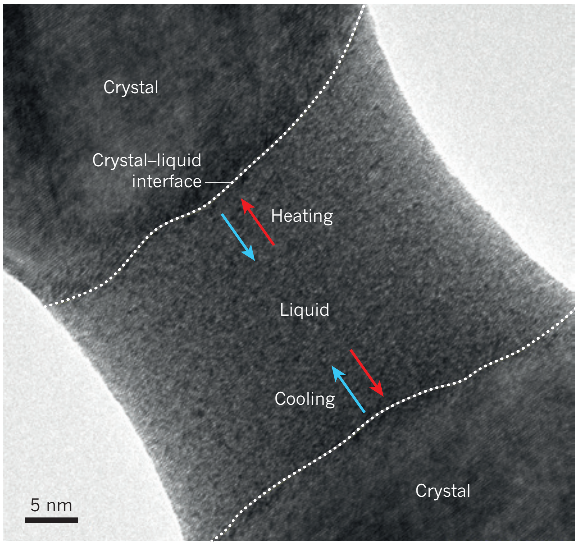
Metallic glasses are pursued for commercial applications because they exhibit attractive mechanical properties such as high strength, elasticity and processability[8]. The advent of metallic-glass formation, along with methods that allow the liquid state of metals to be studied at slow, experimentally accessible timescales, have also been exciting for fundamental science. These developments have enabled study of the properties of metallic liquids, and investigation of both the transition to the crystalline state and the glass transition. However, the fact that multicomponent alloys have been needed for glass formation has complicated the study of metallic glasses.

In multicomponent systems, glass formation depends on atomic-size differences and attraction between atoms of different elements. Glass formation is also affected by the fact that crystallization in alloys typically requires a change in atomic composition: long-range diffusion is needed to establish the difference in composition between the liquid and the growing crystalline phase. Such diffusion has a long timescale and slows down crystallization, facilitating glass formation. But it also obscures the fundamental and ubiquitous aspects of vitrification that would be observed in simple systems. Mao and colleagues' breakthrough allows glass formation to be studied in its purest form, and their findings confirm theoretical and modelling predictions that glass formation can occur in pure metals.

The researchers studied metals in which atoms adopt a 'body-centred cubic' (bcc) arrangement in the crystalline solid phase. But what would happen for metals that adopt different crystal structures, such as the common face-centred cubic (fcc) arrangement? Glass formation is limited only by crystal growth in Mao and co-workers' heating device, and crystal growth rates are slower for bcc crystals than for fcc ones. The R_c values for pure fcc metals are therefore expected to be even higher than those reported by Mao *et al.* for pure bcc metals.

In its most general form, crystallization involves nucleation — the initial formation of tiny crystals called nuclei — and growth. Glass formation competes with the combination of both processes. However, crystallization proceeds through growth into the undercooled liquid phase of the crystal–liquid interface in Mao and co-workers' experiments (Fig. 1). Crystal growth therefore does not depend on nucleation in their system, which means that the R_c values reported by the authors are probably an overestimate for the most general form of vitrification in pure bcc metals. To involve nucleation, direct contact of the liquid phase to a crystalline boundary has to be avoided. Such an experimental realization would allow study of the earliest stages of nucleation — one of the great mysteries of physics.

Experimental investigations of glass formation have typically been carried out on large samples of more than 10^8 atoms, and at long timescales greater than 1 microsecond. By contrast, molecular-dynamics simulations have been limited to small samples of fewer



SCOTT X. MAO

Figure 1 | Ultra-rapid cooling causes a pure metal to form a glass. The micrograph shows a region of molten tantalum between two crystalline regions. Mao and colleagues[1] report that, on ultra-rapid cooling, the crystalline regions grow into the liquid region (blue arrows) until the growth kinetics can no longer keep up with the thermal field defined by the cooling rate. The liquid in front of the interface then 'freezes' into a glass. On heating, the crystal–liquid interface moves out into the crystalline sections of the sample (red arrows).

than 10^5 atoms and short timescales (less than 1 nanosecond), because of the restrictions of available computing power. Our ability to predict experimental results from such simulations has therefore been limited because the properties of metallic glasses are affected by sample size[9] and cooling rates[10]. Mao and colleagues' method now allows us to carry out experiments at spatial and temporal timescales similar to those in simulations. This opens the way to exploring glass formation and its competition with crystallization.

Given that vitrification has previously been observed in ionic melts, aqueous solutions, alloy melts, molecular liquids and polymers, the finding that pure metals can also be glasses suggests that amorphous structures are the most ubiquitous form of condensed matter. ■

Jan Schroers *is in the Department of Mechanical Engineering and Materials Science, Yale University, New Haven, Connecticut 06511, USA.*
e-mail: jan.schroers@yale.edu

1. Zhong, L., Wang, J., Sheng, H., Zhang, Z. & Mao, S. X. *Nature* **512,** 177–180 (2014).
2. Angell, C. A. *Science* **267,** 1924–1935 (1995).
3. Klement, W., Willens, R. H. & Duwez, P. *Nature* **187,** 869–870 (1960).
4. Inoue, A. *Acta Mater.* **48,** 279–306 (2000).
5. Greer, A. L. *Nature* **366,** 303–304 (1993).
6. Turnbull, D. *Contemp. Phys.* **10,** 473–488 (1969).
7. Bhat, M. H. *et al. Nature* **448,** 787–790 (2007).
8. Schroers, J. *Phys. Today* **66,** 32–37 (2013).
9. Volkert, C. A., Donohue, A. & Spaepen, F. *J. Appl. Phys.* **103,** 083539 (2008).
10. Kumar, G., Neibecker, P., Liu, Y.-H. & Schroers, J. *Nature Commun.* **4,** 1536 (2013).

This article was published online on 6 August 2014.

CANCER

One cell at a time

Single-cell DNA sequencing of two breast-cancer types has shown extensive mutational variation in individual tumours, confirming that generation of genetic diversity may be inherent in how tumours evolve. SEE ARTICLE P.155

EDWARD J. FOX & LAWRENCE A. LOEB

Next-generation DNA sequencing has revolutionized the field of cancer genomics[1]. Although this sequencing can identify the most frequent mutation in a population of cells, it struggles to resolve the mutational diversity and multiple genomes of the individual cells that comprise a tumour. Achieving DNA sequencing down to the resolution of a single cell has been a long-held dream for understanding the cellular heterogeneity that is inherent in many complex biological systems and, in particular, for delineating the mixture of genomes in human cancers[2]. On page 155 of this issue, Wang *et al.*[3] report an innovative sequencing method, termed nuc-seq, that achieves almost complete sequencing of whole genomes in single cells.

As a cell prepares to divide, it replicates the DNA in its nucleus. By sorting and sequencing only the newly 'doubled' nuclei, nuc-seq takes advantage of this duplication to achieve lower rates of sequencing errors than most previous techniques[4]. The authors validated their method using targeted duplex sequencing, a protocol that sequences both strands of DNA to identify mutations at exceptionally high accuracy[5]. They suggest that the use of nuc-seq to sequence single-cell genomes, with validation by targeted deep sequencing, will be instrumental in defining the genomic heterogeneity of cancers.

To demonstrate this, Wang *et al.* used their technique to sequence the genomes of multiple single cells from two types of human breast cancer, and found that no two individual tumour cells were genetically identical. As well as the large numbers of mutations that are common to the majority of cells in a tumour, the authors uncovered an even greater number of subclonal and *de novo* mutations (those that are unique to individual cells). They also present estimates, derived from mathematical models, of mutation rates of single cells within tumours. On the basis of these models, they show that distinct types of DNA alteration seem to accumulate at different rates in different tumours, and suggest that two separate 'mutational clocks' operate in cancer. Large-scale, structural changes in DNA (such as amplification and deletion of large blocks of DNA) probably occur early in tumour development, in punctuated bursts of evolution, whereas point mutations may accumulate more gradually, generating extensive subclonal diversity. The authors' findings indicate that slower-growing 'luminal' breast-cancer cells exhibit relatively low mutation rates, whereas cells from clinically more aggressive, 'triple-negative' breast cancers have mutation rates that are 13 times greater than in normal cells.

Nuc-seq and comparable single-cell sequencing methods[6–9] will allow a more detailed understanding of mutational heterogeneity in individual tumours, and will influence our understanding of how cancers evolve and our approach to their treatment. In particular, mutational diversity within a tumour is likely to be predictive of whether resistance to a particular chemotherapy will emerge during treatment, because mutations in genes that render cells resistant to specific drugs may exist before initiation of therapy. This has previously been documented for the failure of certain molecularly tailored cancer treatments[10]. Such findings also reinforce the fact that single, bulk sampling of a tumour — a strategy that is commonly used to select targeted therapies — is not representative of the tumour as a whole.

The total number of mutations that a tumour genome carries, including those present in only a small subset of cells, may in fact underlie the aggressiveness of different cancer subtypes. For example, the extent of genetic diversity within a tumour, and its divergence from normal tissue, probably influences the ability of the immune system to distinguish malignant cells from normal cells. Identifying the mechanisms by which cancer cells generate mutational heterogeneity may therefore

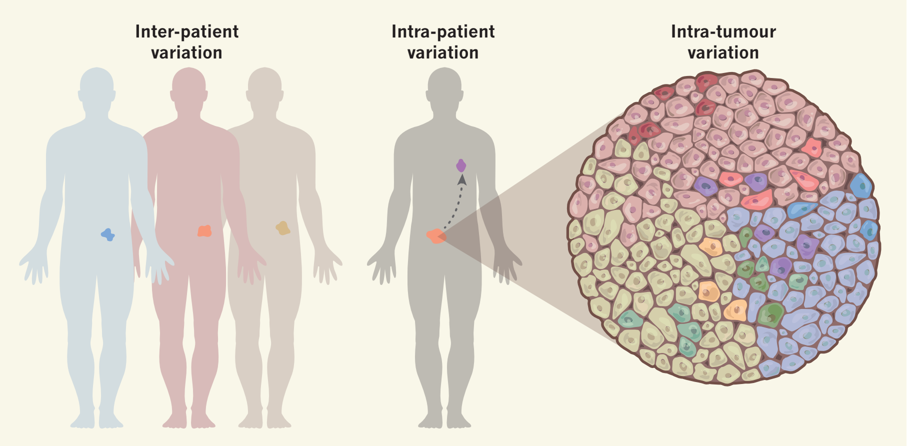


Figure 1 | Levels of diversity. The genetic characteristics of cancers vary between patients, between primary and metastatic tumours in a single patient, and between the individual cells of a tumour. Wang *et al.*[3] present a single-cell, whole-genome sequencing technique that will allow a better understanding of genetic heterogeneity within individual tumours.

present previously unexplored therapeutic targets.

An array of techniques to analyse individual cells has now been developed. It remains to be seen, however, just how robust nuc-seq and other single-cell genomics techniques, such as MALBAC[6], will prove to be. For example, many cancer cells are aneuploid (they carry abnormal numbers of chromosomes), and the application of nuc-seq may be restricted to cancers that do not exhibit aneuploidy. Also, although the cost of genome sequencing continues to decline (albeit more slowly now than in the past), the cost of single-cell genomics and the complexities of the bioinformatic analyses involved are still formidable.

In our quest to decipher cancer genomes, the advent of single-cell sequencing marks a technical milestone. It crystallizes the concept that the genome of each tumour is dynamic and highly diverse, whether we are comparing cancer genomes between tumours of different patients, between anatomically distinct regions of a tumour within a patient or even between individual cells within the same tumour (Fig. 1). Single-cell sequencing will allow us to detect rare mutant subpopulations hidden within cancers that could expand and lead to drug resistance, and thus to avoid unnecessary and potentially harmful administration of ineffective, toxic therapies. Ultimately, the exceptional plasticity of the tumour genome may well prove to be a key characteristic of cancer[11] and a major, as yet untapped, therapeutic vulnerability. ■

Edward J. Fox *and* **Lawrence A. Loeb** *are in the Department of Pathology, University of Washington, Seattle, Washington 98195-7750, USA. L.A.L. is also in the Department of Biochemistry, University of Washington.*
e-mails: eddiefox@uw.edu; laloeb@uw.edu

1. Stratton, M. R., Campbell, P. J. & Futreal, P. A. *Nature* **458,** 719–724 (2009).
2. *Nature Meth.* **11,** 1 (2014).
3. Wang, Y. *et al. Nature* **512,** 155–160 (2014).
4. Navin, N. *et al. Nature* **472,** 90–94 (2011).
5. Schmitt, M. W. *et al. Proc. Natl Acad. Sci. USA* **109,** 14508–14513 (2012).
6. Zong, C., Lu, S., Chapman, A. R. & Xie, X. S. *Science* **338,** 1622–1626 (2012).
7. Shapiro, E., Biezuner, T. & Linnarsson, S. *Nature Rev. Genet.* **14,** 618–630 (2013).
8. Xu, X. *et al. Cell* **148,** 886–895 (2012).
9. Hou, Y. *et al. Cell* **148,** 873–885 (2012).
10. Tougeron, D. *et al. Ann. Oncol.* **24,** 1267–1273 (2013).
11. Loeb, L. A., Springgate, C. F. & Battula, N. *Cancer Res.* **34,** 2311–2321 (1974).

This article was published online on 30 July 2014.

ASTRONOMICAL INSTRUMENTATION

Atmospheric blurring has a new enemy

A fully automated optics system that corrects atmospheric blurring of celestial objects has imaged 715 star systems thought to harbour planets, completing each observation in less time than it takes to read this article.

BRENT ELLERBROEK

At the time of writing, observations from the Kepler Space Mission have yielded more than 975 confirmed exoplanet detections from 4,234 candidates[1]. These candidates are identified by small, periodic drops in the brightness of the star, indicating that a planet might be transiting in front of it[2]. This is perhaps the conceptually simplest method of finding exoplanets, and it remains the only approach that can find planets with the proper orbit and radius to potentially support life. However, follow-up observations of high spatial resolution are needed to confirm and characterize each candidate system detected by the Kepler mission. Writing in *The Astrophysical Journal*, Law *et al.*[3] describe how they have used a robotic adaptive optics system[4,5] to follow up 715 of the Kepler candidate star systems in just 36 hours of observing time.

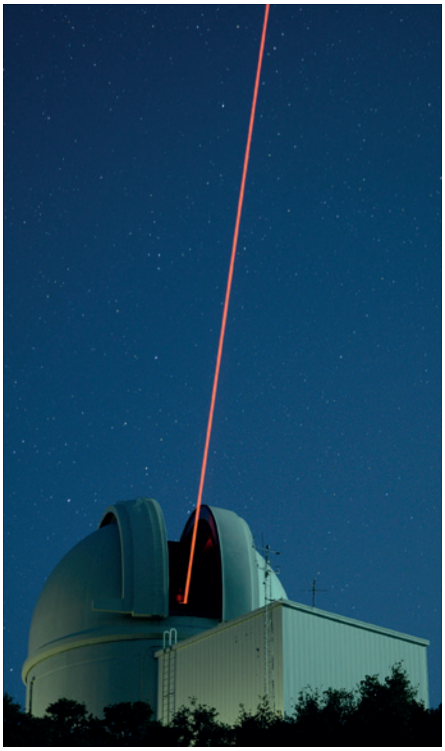
In planetary-transit observations, the size of the planet can be inferred from the relative dip in star brightness measured during the transit. Only a small fraction of exoplanets will transit their star when viewed from Earth, but a statistical analysis[6] of the Kepler candidates detected in observations of more than 100,000 stars indicates that exoplanets may be relatively common. This includes Earth-sized planets with orbits that would permit liquid water to exist on the planets' surfaces[6]. Because Kepler was designed to continually monitor many thousands of stars, its images lack the spatial resolution needed to characterize individual star systems in further detail. This means that various false-positive detections — for example, those associated with the partial eclipse of a star in a binary system by its companion star — cannot be ruled out, and that possible binary host stars (or stars in the foreground or background by coincidence) cannot necessarily be identified.

The presence or absence of a stellar companion to a 'primary host star' is important information for understanding the formation and development of planetary systems. It also affects the estimation of the planet's size from the transit: the relationship between star brightness and planetary radius is more complex if there is a stellar companion, leading to incorrect results if the existence of the companion star is unknown. For these and other reasons, follow-up observations with high angular resolution are needed to fully understand each Kepler candidate.

High-resolution follow-up images could be collected by a space-based observatory such as the Hubble Space Telescope, but observing thousands of candidate systems would monopolize this limited resource. Obtaining such images from the ground is made difficult by the blurring ('seeing') introduced by atmospheric turbulence, and by the resulting inhomogeneity in the density and refractive index of the air. In the past two decades, ground-based observatories have begun using a technology known as adaptive optics to measure and correct this blurring in real time[7]. Many of these systems now use lasers to create artificial 'guide stars' on the sky to measure the blurring, and then correct it for science targets that are themselves too faint to be used for such measurement — as is the case for many of the Kepler candidates. Adaptive optics surveys of the candidate systems began in 2011–12 (refs 8,9), but these initial studies were limited to fewer than 100 targets because of the time taken to set up and initiate each observation, typically at least 15–20 minutes per target for most current adaptive optics systems.

The robotic adaptive optics system (Robo-AO) used by Law and colleagues supersedes these constraints. The system has been designed[4,5] for highly efficient, automated high-resolution observing on 1- to 3-metre-class telescopes, and has been mounted on the 60-inch (1.5-metre) telescope at the Palomar Observatory in California (Fig. 1). The atmospheric blurring at Palomar Observatory is typically about 0.65 arcseconds. Robo-AO sharpens star images to about 0.12–0.15 arcseconds in diameter[4,5] — not far from the 0.09-arcsecond value that is theoretically possible with a 1.5-metre telescope in space. This performance has enabled Law *et al.* to resolve 53 of the 715 Kepler candidates observed by Robo-AO so far into multiple stars. Forty-three of these 53 are new discoveries, including one that is a probable false positive for a candidate exoplanet.

Of course, automated observing at a rate of 200–250 targets per night, as Law and co-workers have done, creates a substantial data cleaning and analysis task. To detect and characterize companion stars that are significantly fainter than their primaries, the authors have developed a fully



CHRISTOPH BARANEC

Figure 1 | Creating artificial guide stars. The robotic adaptive optics system (Robo-AO)[4,5], which has been used by Law and colleagues[3] to observe star systems that are thought to host planets, projects a laser beam above the Palomar 60-inch telescope to generate an artificial guide star. This is then used to sense and correct atmospheric blurring. The ultraviolet beam is not visible to the human eye, but it can be seen in digital cameras after their internal filters are removed.

automated data-processing pipeline. More than 800 short (115-millisecond) exposures are collected of each target, and these must first be calibrated, centred and averaged into a single image. The light from the primary star is then subtracted from the image, and an automated target-detection algorithm is applied. Companion stars that are as faint as one one-hundredth to four one-thousandths of the primary can be detected in median to good atmospheric conditions; this is not faint enough to find exoplanets, but more than sufficient to identify many companion stars. Once the brightness of a companion star has been measured, its mass and diameter can be determined from standard stellar models and from the characteristics of the primary star.

Law *et al.* have provided updated estimates for the planetary radii of each of the Kepler candidates with a fainter companion star. Five small planet candidates have been confirmed to be less than twice the diameter of Earth, but a larger number of other candidates could be significantly bigger than this threshold if they are found to be orbiting the fainter companion star instead of the primary (this would require future observations of their transits with Robo-AO or some other high-resolution system). The team also suggests that several of the stars with multiple Kepler candidate planets are likely to be coincident multiples — two separate planetary systems orbiting both stars of the binary pair. In addition, the Robo-AO observations so far yield plausible (98% confidence) evidence[3] that giant planets with orbital periods of less than 15 days are two to three times more likely than longer-period planets to be found in binary-star systems. This suggests that companion stars have a role in creating close-in giant planets and stabilizing their orbits. The researchers expect to observe every Kepler candidate using Robo-AO by the end of 2014 to confirm this conclusion, and aim to develop a more comprehensive statistical sample of planetary systems associated with binary stars.

More generally, these results are a convincing indication that laser-guide-star adaptive optics is now ready for highly efficient, quantitatively precise, high-resolution astronomical observations. Current and future adaptive optics systems on much larger telescopes than the Palomar 60-inch telescope — such as Keck, Gemini and the European Southern Observatory's Very Large Telescope — can produce even sharper images of even fainter objects, but much work will be needed to match the degree of automation and efficiency already demonstrated by Robo-AO. ■

Brent Ellerbroek *is in the Instrumentation Department, Thirty Meter Telescope Observatory Corporation, Pasadena, California 90807, USA.*
e-mail: brente@caltech.edu

1. http://kepler.nasa.gov
2. Fressin, F. *et al. Nature* **482,** 195–198 (2012).
3. Law, N. M. *et al. Astrophys. J.* **791,** 35 (2014).
4. Baranec, C. *et al. J. Vis. Exp.* **72,** e50021 (2013).
5. Baranec, C. *et al. Astrophys. J. Lett.* **790,** L8 (2014).
6. Petigura, E. A., Howard, A. W. & Marcy, G. W. *Proc. Natl Acad. Sci. USA* **110,** 19273–19278 (2013).
7. Davies, R. & Kasper, M. *Annu. Rev. Astron. Astrophys.* **50,** 305–351 (2012).
8. Adams, E. R. *et al. Astron. J.* **144,** 42 (2012).
9. Lillo-Box, J., Barrado, D. & Bouy, H. *Astron. Astrophys.* **546,** A10 (2012).

STRUCTURAL BIOLOGY

Corralling a protein–degradation regulator

The crystal structure of the COP9 signalosome, a large protein complex that regulates intracellular protein degradation, reveals how the complex achieves exquisite specificity for its substrates. SEE ARTICLE P.161

RAYMOND J. DESHAIES

Some 20 years ago[1], an enzyme complex was linked to the dramatic changes in development that occur when seedlings push through the soil and encounter sunlight. This complex, named the COP9 signalosome (CSN), is now thought to be common to all animals, plants and fungi. The CSN is involved in protein degradation, but because of its complicated structure, detailed knowledge of how its activity is controlled has remained elusive. On page 161 of this issue, Lingaraju *et al.*[2] report the crystallization of the CSN and determine its structure to a resolution of a remarkable 3.8 ångströms.

The CSN consists of eight protein subunits, CSN1–8, and regulates a family of enzyme complexes called cullin–RING E3 ubiquitin ligases (CRLs)[3], which modify their target proteins by attaching ubiquitin proteins to them. Ubiquitin modifications can have many effects on proteins, from influencing their cellular location to causing their degradation. In fact, the cullin protein that makes up the backbone of each CRL must itself be modified by a ubiquitin-like protein, NEDD8, before it can function as a ubiquitin ligase. The CSN inhibits this activity by detaching NEDD8 from cullin, and can also bind 'deneddylated' CRLs, thereby maintaining CRL inactivity after NEDD8 removal[4–7].

The CSN structure described by Lingaraju and colleagues brings to mind a widely splayed hand on which a small box sits askew, topped by a tomato (Fig. 1). Like a hand, the CSN has five digits (the amino-terminal ends of CSN1, 2, 4, 7, and 3 plus 8) projecting from an organizing centre, the palm. The palm is formed by the 'winged-helix' subdomains of these subunits, which associate to form a horseshoe-shaped structure. Resting on the hand is the box, formed by bundling of the carboxy-terminal ends of each subunit. Sitting atop this platform is the CSN5–CSN6 tomato.

Whereas some aspects of the CSN structure were anticipated from previous work on related proteins, it is a big surprise that the structure obtained by Lingaraju and co-workers is in an inactive configuration. The active site of the CSN is specified by a 'JAMM' domain in the CSN5

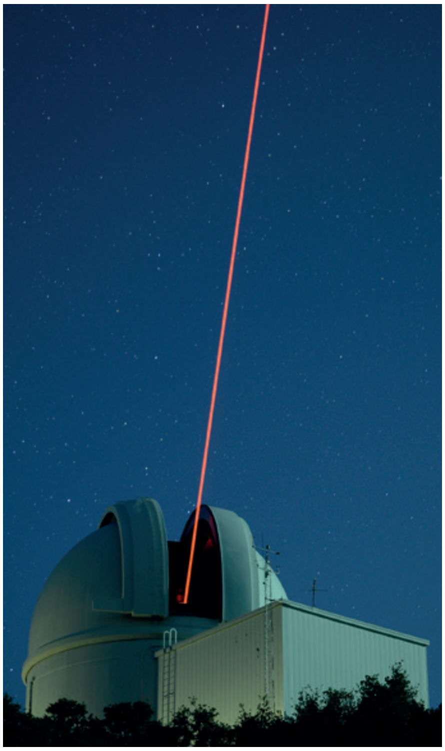

CHRISTOPH BARANEC

Figure 1 | Creating artificial guide stars. The robotic adaptive optics system (Robo-AO)[4,5], which has been used by Law and colleagues[3] to observe star systems that are thought to host planets, projects a laser beam above the Palomar 60-inch telescope to generate an artificial guide star. This is then used to sense and correct atmospheric blurring. The ultraviolet beam is not visible to the human eye, but it can be seen in digital cameras after their internal filters are removed.

automated data-processing pipeline. More than 800 short (115-millisecond) exposures are collected of each target, and these must first be calibrated, centred and averaged into a single image. The light from the primary star is then subtracted from the image, and an automated target-detection algorithm is applied. Companion stars that are as faint as one one-hundredth to four one-thousandths of the primary can be detected in median to good atmospheric conditions; this is not faint enough to find exoplanets, but more than sufficient to identify many companion stars. Once the brightness of a companion star has been measured, its mass and diameter can be determined from standard stellar models and from the characteristics of the primary star.

Law *et al.* have provided updated estimates for the planetary radii of each of the Kepler candidates with a fainter companion star. Five small planet candidates have been confirmed to be less than twice the diameter of Earth, but a larger number of other candidates could be significantly bigger than this threshold if they are found to be orbiting the fainter companion star instead of the primary (this would require future observations of their transits with Robo-AO or some other high-resolution system). The team also suggests that several of the stars with multiple Kepler candidate planets are likely to be coincident multiples — two separate planetary systems orbiting both stars of the binary pair. In addition, the Robo-AO observations so far yield plausible (98% confidence) evidence[3] that giant planets with orbital periods of less than 15 days are two to three times more likely than longer-period planets to be found in binary-star systems. This suggests that companion stars have a role in creating close-in giant planets and stabilizing their orbits. The researchers expect to observe every Kepler candidate using Robo-AO by the end of 2014 to confirm this conclusion, and aim to develop a more comprehensive statistical sample of planetary systems associated with binary stars.

More generally, these results are a convincing indication that laser-guide-star adaptive optics is now ready for highly efficient, quantitatively precise, high-resolution astronomical observations. Current and future adaptive optics systems on much larger telescopes than the Palomar 60-inch telescope — such as Keck, Gemini and the European Southern Observatory's Very Large Telescope — can produce even sharper images of even fainter objects, but much work will be needed to match the degree of automation and efficiency already demonstrated by Robo-AO. ■

Brent Ellerbroek *is in the Instrumentation Department, Thirty Meter Telescope Observatory Corporation, Pasadena, California 90807, USA.*
e-mail: brente@caltech.edu

1. http://kepler.nasa.gov
2. Fressin, F. *et al. Nature* **482,** 195–198 (2012).
3. Law, N. M. *et al. Astrophys. J.* **791,** 35 (2014).
4. Baranec, C. *et al. J. Vis. Exp.* **72,** e50021 (2013).
5. Baranec, C. *et al. Astrophys. J. Lett.* **790,** L8 (2014).
6. Petigura, E. A., Howard, A. W. & Marcy, G. W. *Proc. Natl Acad. Sci. USA* **110,** 19273–19278 (2013).
7. Davies, R. & Kasper, M. *Annu. Rev. Astron. Astrophys.* **50,** 305–351 (2012).
8. Adams, E. R. *et al. Astron. J.* **144,** 42 (2012).
9. Lillo-Box, J., Barrado, D. & Bouy, H. *Astron. Astrophys.* **546,** A10 (2012).

STRUCTURAL BIOLOGY

Corralling a protein–degradation regulator

The crystal structure of the COP9 signalosome, a large protein complex that regulates intracellular protein degradation, reveals how the complex achieves exquisite specificity for its substrates. SEE ARTICLE P.161

RAYMOND J. DESHAIES

Some 20 years ago[1], an enzyme complex was linked to the dramatic changes in development that occur when seedlings push through the soil and encounter sunlight. This complex, named the COP9 signalosome (CSN), is now thought to be common to all animals, plants and fungi. The CSN is involved in protein degradation, but because of its complicated structure, detailed knowledge of how its activity is controlled has remained elusive. On page 161 of this issue, Lingaraju *et al.*[2] report the crystallization of the CSN and determine its structure to a resolution of a remarkable 3.8 ångströms.

The CSN consists of eight protein subunits, CSN1–8, and regulates a family of enzyme complexes called cullin–RING E3 ubiquitin ligases (CRLs)[3], which modify their target proteins by attaching ubiquitin proteins to them. Ubiquitin modifications can have many effects on proteins, from influencing their cellular location to causing their degradation. In fact, the cullin protein that makes up the backbone of each CRL must itself be modified by a ubiquitin-like protein, NEDD8, before it can function as a ubiquitin ligase. The CSN inhibits this activity by detaching NEDD8 from cullin, and can also bind 'deneddylated' CRLs, thereby maintaining CRL inactivity after NEDD8 removal[4–7].

The CSN structure described by Lingaraju and colleagues brings to mind a widely splayed hand on which a small box sits askew, topped by a tomato (Fig. 1). Like a hand, the CSN has five digits (the amino-terminal ends of CSN1, 2, 4, 7, and 3 plus 8) projecting from an organizing centre, the palm. The palm is formed by the 'winged-helix' subdomains of these subunits, which associate to form a horseshoe-shaped structure. Resting on the hand is the box, formed by bundling of the carboxy-terminal ends of each subunit. Sitting atop this platform is the CSN5–CSN6 tomato.

Whereas some aspects of the CSN structure were anticipated from previous work on related proteins, it is a big surprise that the structure obtained by Lingaraju and co-workers is in an inactive configuration. The active site of the CSN is specified by a 'JAMM' domain in the CSN5 subunit. Typically, the active sites of the enzymes in the JAMM family contain a zinc ion (Zn^{2+}) bound by three evolutionarily conserved amino-acid residues (two histidines and an aspartate), with the remaining ligand-binding site of Zn^{2+}occupied by a water molecule that has been activated by another evolutionarily conserved amino acid, glutamate 76 (Glu 76; ref. 8). This activated water molecule detaches ubiquitin or ubiquitin-like proteins from their targets by hydrolysis. Whereas the histidine and aspartate residues of CSN5 are positioned as expected in the CSN structure, the water molecule is replaced by another amino acid, Glu 104. This explains a long-standing puzzle: whereas other JAMM-containing proteins efficiently cleave model substrates, such as ubiquitin with a rhodamine dye attached to its C terminus, purified CSN does so only poorly.

Lingaraju *et al.* tested the role of Glu 104 in CSN regulation by performing enzyme assays on CSN complexes in which Glu 104 was mutated. This mutant cleaved ubiquitin–rhodamine much faster than the natural enzyme, indicating that Glu104–Zn^{2+} binding might keep the CSN in an inactive state when it is free from CRL. Notably, mutation of the adjacent residue, threonine 103, results in defective development of the nervous system in fruit flies[9], which suggests that Glu 104-mediated regulation is required for proper control of CSN activity *in vivo*.

The inhibited state of unbound CSN raises the obvious question of how the CSN gains its activity on binding CRLs. The authors used computer-modelling studies to compare their crystal structure of free CSN with a structure determined by electron microscopy[7] in which the CSN was bound to a CRL enzyme to which NEDD8 is attached. This comparison showed clearly that, to reconcile the two structures, substantial rearrangements of CSN2, CSN4 and CSN5–CSN6 must occur when the CSN and CRL bind (Fig. 1). In particular, movements in CSN4 and CSN6 must lead to a change in the CSN4–CSN6 interface.

To probe the significance of this interface, Lingaraju and colleagues deleted a β-hairpin loop in CSN6 that contributes to its interaction with CSN4. Surprisingly, the resulting complex, like the Glu 104 mutant, efficiently cleaved ubiquitin–rhodamine. It also deneddylated CRL more than four times faster than did the wild-type enzyme. These observations make it tempting to speculate that CSN4 is the signalosome's CRL sensor, and that CSN4 movement during CRL binding triggers a cascade of rearrangements transmitted through CSN6 that prise CSN5's Glu 104 residue away from Zn^{2+}, so that Glu 76 can move into position, activating the CSN. However, when the authors made a double mutant lacking both the CSN6 loop and Glu 104, they found it to be more active than either individual mutant, suggesting that these two mutations have independent effects, rather than acting in a linear cascade. Furthermore, the N-terminal region of CSN4 does not seem to make strong contact with CRL[7], indicating that the CSN's CRL sensor may be in another subunit.

This study highlights a crucial lesson on the use of evolutionary conservation to predict enzyme regulation. Comparing the crystal structure of the CSN with those of the JAMM-containing enzymes AMSH-LP (ref. 10) and Rpn11 (refs 11, 12) reveals that, although all three use the same amino acids to coordinate Zn^{2+} and the activated water molecule, their activities are controlled in markedly different ways. AMSH-LP seems to be constitutively active, Rpn11 activity is promoted by rearrangements that bring the enzyme and its target substrate into proximity[13,14], and CSN5 is activated by substrate-driven relief of inhibition. Strikingly, CSN5 is inhibited by distinct mechanisms depending on whether the subunit is on its own[15] or integrated into the CSN. Although some generalizations apply across the JAMM family, it is clear that each member has its own distinctive features.

What lies ahead for research on the CSN? It will be fascinating to examine the structure of different CSN mutants, to work out the mechanism by which binding to CRLs brings about major conformational changes. It would also be wonderful to see a CSN–NEDD8–CRL complex in its full glory, to gain an atomic-level view of the CSN–CRL interface and how it might be influenced by NEDD8 or substrates that bind to CRL. Another question is whether binding of the CSN to neddylated or deneddylated CRL promotes the same conformational change in the CSN.

Binding and kinetic studies of the CSN and the mutated complexes reported by Lingaraju *et al.* should reveal whether the CSN's catalytic rate is determined by the conformational rearrangement that occurs on CRL binding. Furthermore, *in vivo* studies with Glu 104 and CSN6-loop mutants should show why free CSN must be inhibited.

Finally, this structure may help the design of drugs that act on the CSN, which could be an attractive target for the treatment of breast and liver cancer[16,17]. Although detailed characterization of CSN inhibitors has not been reported, my laboratory has identified several candidates through high-throughput screening (PubChem AID652009). The surprising observations reported by Lingaraju *et al.* suggest that it may be possible to inhibit the CSN but spare other JAMM proteins, by interfering with the active-site rearrangement that occurs when the CSN and CRL bind. ■

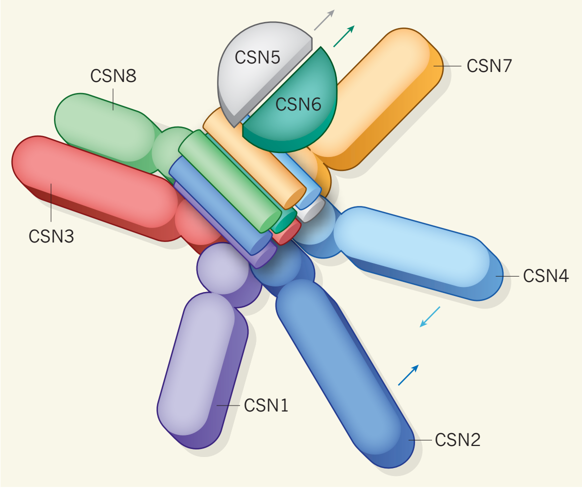


Figure 1 | Structure of the COP9 signalosome (CSN). This enzyme complex is comprised of eight CSN protein subunits. Six subunits make up the base of the CSN, a splayed 'hand' in which the proteins' N-terminal ends are at the fingertips and their winged-helix domains, drawn as circles, assemble to form the palm (partially obscured). The C-terminal ends of each protein are bundled together into a 'box' that sits askew on the hand. The CSN5 and CSN6 subunits associate intimately to form a 'tomato' sitting on the box. Lingaraju *et al.*[2] report that the CSN is inactive until it binds to its target, a cullin–RING E3 ubiquitin-ligase enzyme complex. On binding, the CSN undergoes activating conformational changes, indicated by coloured arrows that represent the movements of the altered subunits. For simplicity, the box is drawn as a uniform bundle, and so does not represent the actual position and length of each C terminus. (Figure adapted from Fig. 1 of ref. 2.)

Raymond J. Deshaies *is in the Division of Biology and Biological Engineering, California Institute of Technology, Pasadena, California 91125, USA.*
e-mail: deshaies@caltech.edu

1. Wei, N., Chamovitz, D. A. & Deng, X.-W. *Cell* **78,** 117–124 (1994).
2. Lingaraju, G. M. *et al. Nature* **512,** 161–165 (2014).
3. Lyapina, S. *et al. Science* **292,** 1382–1385 (2001).
4. Cope, G. A. *et al. Science* **298,** 608–611 (2002).
5. Fischer, E. S. *et al. Cell* **147,** 1024–1039 (2011).
6. Emberley, E. D., Mosadeghi, R. & Deshaies, R. J. *J. Biol. Chem.* **287,** 29679–29689 (2012).
7. Enchev, R. I. *et al. Cell Rep.* **2,** 616–627 (2012).
8. Ambroggio, X. I., Rees, D. C. & Deshaies, R. J. *PLoS Biol.* **2,** e2 (2004).
9. Suh, G. S. B. *et al. Neuron* **33,** 35–46 (2002).
10. Sato, Y. *et al. Nature* **455,** 358–362 (2008).
11. Pathare, G. R. *et al. Proc. Natl Acad. Sci. USA* **111,** 2984–2989 (2014).
12. Worden, E. J., Padovani, C. & Martin, A. *Nature Struct. Mol. Biol.* **21,** 220–227 (2014).
13. Śledź, P. *et al. Proc. Natl Acad. Sci. USA* **110,** 7264–7269 (2013).
14. Matyskiela, M. E., Lander, G. C. & Martin, A. *Nature Struct. Mol. Biol.* **20,** 781–788 (2013).
15. Echalier, A. *et al. Proc. Natl Acad. Sci. USA* **110,** 1273–1278 (2013).
16. Adler, A. S. *et al. Cancer Res.* **68,** 506–515 (2008).
17. Lee, Y.-H. *et al. Oncogene* **30,** 4175–4184 (2011).

This article was published online on 16 July 2014.

doi:10.1038/nature13570

Limits on fundamental limits to computation

Igor L. Markov[1]*

An indispensable part of our personal and working lives, computing has also become essential to industries and governments. Steady improvements in computer hardware have been supported by periodic doubling of transistor densities in integrated circuits over the past fifty years. Such Moore scaling now requires ever-increasing efforts, stimulating research in alternative hardware and stirring controversy. To help evaluate emerging technologies and increase our understanding of integrated-circuit scaling, here I review fundamental limits to computation in the areas of manufacturing, energy, physical space, design and verification effort, and algorithms. To outline what is achievable in principle and in practice, I recapitulate how some limits were circumvented, and compare loose and tight limits. Engineering difficulties encountered by emerging technologies may indicate yet unknown limits.

Emerging technologies for computing promise to outperform conventional integrated circuits in computation bandwidth or speed, power consumption, manufacturing cost, or form factor[1,2]. However, razor-sharp focus on any one nascent technology and its benefits sometimes neglects serious limitations or discounts ongoing improvements in established approaches. To foster a richer context for evaluating emerging technologies, here I review limiting factors and the salient trends in computing that determine what is achievable in principle and in practice. Several fundamental limits remain substantially loose, possibly indicating viable opportunities for emerging technologies. To clarify this uncertainty, I examine the limits on fundamental limits.

Universal and general-purpose computers

If we view clocks and watches as early computers, it is easy to see the importance of long-running calculations that can be repeated with high accuracy by mass-produced devices. The significance of programmable digital computers became clear at least 200 years ago, as illustrated by Jacquard looms in textile manufacturing. However, the existence of universal computers that can efficiently simulate (almost) all other computing devices—analogue or digital—was only articulated in the 1930s by Church and Turing (Turing excluded quantum physics when considering universality)[3]. Efficiency was studied from a theoretical perspective at first, but strong demand in military applications in the 1940s led Turing and von Neumann to develop detailed hardware architectures for universal computers—Turing's design (Pilot ACE) was more efficient, but von Neumann's was easier to program. The stored-program architecture made universal computers practical in the sense that a single computer design could be effective in many diverse applications if supplied with appropriate software. Such practical universality thrives (1) in economies of scale in computer hardware and (2) among extensive software stacks. Not surprisingly, the most sophisticated and commercially successful computer designs and components, such as Intel and IBM central processing units (CPUs), were based on the von Neumann paradigm. The numerous uses and large markets of general-purpose chips, as well as the exact reproducibility of their results, justify the enormous capital investment in the design, verification and manufacturing of leading-edge integrated circuits. Today general-purpose CPUs power cloud server-farms and displace specialized (but still universal) mainframe processors in many supercomputers. Emerging universal computers based on field-programmable gate-arrays and general-purpose graphics processing units outperform CPUs in some cases, but their efficiencies remain complementary to those of CPUs. The success of deterministic general-purpose computing is manifest in the convergence of diverse functionalities in portable, inexpensive smartphones. After steady improvement, general-purpose computing displaced entire industries (newspapers, photography, and so on) and launched new applications (video conferencing, GPS navigation, online shopping, networked entertainment, and so on)[4]. Application-specific integrated circuits streamline input–output and networking, or optimize functionalities previously performed by general-purpose hardware. They speed up biomolecular simulation 100-fold[5,6] and improve the efficiency of video decoding 500-fold[7], but they require design efforts with a keen understanding of specific computations, impose high costs and financial risks, need markets where general-purpose computers lag behind, and often cannot adapt to new algorithms. Recent techniques for customizable domain-specific computing[8] offer better tradeoffs, while many applications favour the combination of general-purpose hardware and domain-specific software, including specialized programming languages[9,10] such as Erlang, which was used to implement the popular Whatsapp instant messenger.

Limits as aids to evaluating emerging technologies

Without sufficient history, we cannot extrapolate scaling laws for emerging technologies, yet expectations run high. For example, new proposals for analogue processors appear frequently (as illustrated by adiabatic quantum computers), but fail to address concerns about analogue computing, such as its limitations on scale, reliability, and long-running error-free computation. General-purpose computers meet these requirements with digital integrated circuits and now command the electronics market. In comparison, quantum computers—both digital and analogue—hold promise only in niche applications and do not offer faster general-purpose computing because they are no faster for sorting and other specific tasks[11–13]. In exaggerating the engineering impact of quantum computers, the popular press has missed this important point. But in scientific research, attempts to build quantum computers may help in simulating quantum-chemical phenomena and reveal new fundamental limits. The sections 'Asymptotic space-time limits' and 'Conclusions' below discuss the limits on emerging technologies.

Technology extrapolation versus fundamental limits

The scaling of commercial computing hardware regularly runs into formidable obstacles[1,2], but near-term technological advances often circumvent

[1] EECS Department, The University of Michigan, Ann Arbor, Michigan 48109-2121, USA. *Present address: Google Inc., 1600 Amphitheatre Parkway, Mountain View, California 94043, USA.

Table 1 | Some of the known limits to computation

Limits	Engineering	Design and validation	Energy, time	Space, time	Information, complexity
Fundamental	Abbe (diffraction); Amdahl; Gustafson	Error-correction and dense codes; fault-tolerance thresholds	Einstein ($E = mc^2$); Heisenberg ($\Delta E \Delta t$); Landauer ($kT\ln 2$); Bremermann; adiabatic theorems	Speed of light; Planck scale; Bekenstein; Fisher ($T(n)^{1/(d+1)}$)	Shannon channel capacity; Holevo bound; NC, NP, *#P*; decidability
Material	Dielectric constant; carrier mobility; surface morphology; fabrication-related	Analytical and numerical modelling	Conductivity; permittivity; bandgap; heat flow	Propagation speed; atomic spacing; no gravitational collapse	Information transfer between carriers
Device	Gate dielectric; channel charge control; leakage; latency; cross-talk; ageing	Compact modelling; parameter selection	CMOS; quantum; charge-centric; signal-to-noise ratio; energy conversion	Interfaces and contacts; entropy density; entropy flow; size and delay variation; universality	
Circuit	Delay; inductance; thermal-related; yield; reliability; input–output	Interconnect; test; validation	Dark, darker, dim and grey silicon; interconnect; cooling efficiency; power density; power supply; two or three dimensions		Circuit complexity bounds
System and software	Specification; implementation; validation; cost		Synchronization; physical integration; parallelism; *ab initio* limits (Lloyd)		The 'consistency, availability, partitioning tolerance' (CAP) theorem

Summary of material from refs 5, 13–15, 17, 18, 22, 23, 26, 31, 39, 41, 42, 46, 48–50, 53, 54, 57–60, 62, 63, 65, 74–76, 78, 87, 96, 98 and 99.

them. The ITRS[14] keeps track of such obstacles and possible solutions with a focus on frequently revised consensus estimates. For example, consensus estimates initially predicted 10-GHz CPUs for the 45-nm technology node[15], versus the 3–4-GHz range seen in practice. In 2004, the unrelated Quantum Information Science and Technology Roadmap[16] forecast 50 'digital' physical qubits by 2012. Such optimism arose by assuming technological solutions long before they were developed and validated, and by overlooking important limits. The authors of refs 17 and 18 classify the limits to devices and interconnects as fundamental, material, device, circuit, and system limits. These categories define the rows of Table 1, and the columns reflect the sections of this Review in which I examine the impact of specific limits on feasible computing technologies, looking for 'tight' limits, which obstruct the long-term improvement of key parameters.

Engineering obstacles

Engineering obstacles limit specific technologies and choices. For example, a key bottleneck today is integrated circuit manufacture, which packs billions of transistors and wires in several square centimetres of silicon, with astronomically low defect rates. Layers of material are deposited on silicon and patterned with lasers, fabricating all circuit components simultaneously. Precision optics and photochemical processes ensure accuracy.

Limits on manufacturing

No account of limits to computing is complete without the Abbe diffraction limit: light with wavelength λ, traversing a medium with refractive index η, and converging to a spot with angle θ (perhaps focused by a lens) creates a spot with diameter $d = \lambda/\mathrm{NA}$, where $\mathrm{NA} = \eta \sin\theta$ is the numerical aperture. NA reaches 1.4 for modern optics, so it would seem that semiconductor manufacturing is limited to feature sizes of $\lambda/2.8$. Hence, argon-fluoride lasers with a wavelength of 193 nm should not support photolithographic manufacturing of transistors with 65-nm features. Yet these lasers can support subwavelength lithography even for the 45-nm to 14-nm technology nodes if asymmetric illumination and computational lithography are used[19]. In these techniques, one starts with optical masks that look like the intended image, but when the image gets blurry, the masks are altered by gently shifting the edges to improve the image, possibly eventually giving up the semblance between the original mask and the final image. Clearly, some limits are formulated to be broken! Ten years ago, researchers demonstrated the patterning of nanomaterials by live viruses[20]. Known virions exceed 20 nm in diameter, whereas subwavelength lithography using a 193-nm ArF laser was recently extended to 14-nm semiconductor manufacturing[14]. Hence, viruses and microorganisms are no longer at the forefront of semiconductor manufacturing. Extreme ultraviolet (X-ray) lasers have been energy-limited, but are improving. Their use requires changing the optics from refractive to reflective. Additional progress in multiple patterning and directed self-assembly promises to support photolithography beyond the 10-nm technology node.

Limits on individual interconnects

Despite the doubling of transistor density with Moore's law[21], semiconductor integrated circuits would not work without fast and dense interconnects. Copper wires can be either fast or dense, but not both at the same time—a smaller cross-section increases electrical resistance, while greater height or width increase parasitic capacitance with neighbouring wires (wire delay grows with the product of resistance and capacitance, *RC*). As pointed out in 1995 by an Intel researcher, on-chip interconnect scaling has become the real limiter of high-performance integrated circuits[22]. The scaling of interconnect is also moderated by electron scattering against rough edges of metallic wires[18], which is inevitable with atomic-scale wires. Hence, integrated circuit interconnect stacks have evolved[15,23] from four equal-pitch layers in 2000 to 16 layers with some wires up to 32 times thicker than others (as in Fig. 3) including a large amount of dense (thin) wiring and fast (thick) wires used for global on-chip communication (Fig. 3). Aluminium and copper remain unrivalled for conventional interconnects and can be combined in short wires[98]; carbon-nanotube and spintronic interconnects are also evaluated in ref. 98. Photonic waveguides and radio frequency links offer alternative integrated circuit interconnect[24,25], but still obey fundamental limits derived from Maxwell's equations, such as the maximum propagation speed of electromagnetic waves[18]. The number of input–output links can only grow with the perimeter or surface area of a chip, whereas chip capacity grows with area or volume, respectively.

Limits on conventional transistors

Transistors are limited by their tiniest feature—the width of the gate dielectric—which recently reached the size of several atoms (Fig. 1), creating problems: (1) a few missing atoms can alter transistor performance, (2) manufacturing variation makes all the transistors slightly different (Fig. 2), (3) electric current tends to leak through thin narrow dielectrics[17]. Therefore, transistors are redesigned with wider dielectric layers[26] that surround a fin shape (Fig. 4). Such configurations improve the control of the electric field, reduce current densities and leakage, and diminish process variations. Each field effect transistor (FET) can use several fins, extending transistor scaling by several generations. Semiconductor manufacturers adopted such FinFETs for upcoming technology nodes. Going a step further, in tunnelling transistors[27], a gate wraps around the channel to control the tunnelling rate.

Limits on design effort

In the 1980s, Mead and Conway formalized integrated circuit design using a regular grid, enabling automated layout through algorithms. But the

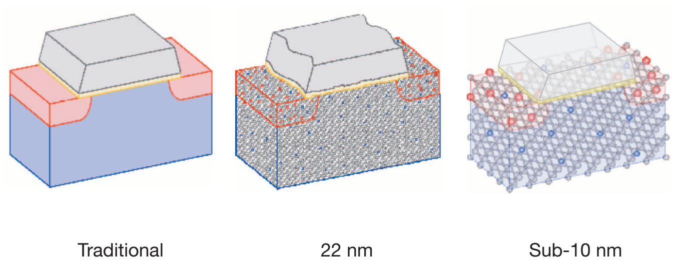


Figure 1 | As a metal oxide–semiconductor field effect transistor (MOSFET) shrinks, the gate dielectric (yellow) thickness approaches several atoms (0.5 nm at the 22-nm technology node). Atomic spacing limits the device density to one device per nanometre, even for radical devices. For advanced transistors, grey spheres indicate silicon atoms, while red and blue spheres indicate dopant atoms (intentional impurities that alter electrical properties). Image redrawn from figure 1 of http://cnx.org/content/m32874/latest/, with permission from Gold Standard Simulations.

resulting optimization problems remain difficult to solve, and heuristics are only good enough for practical use. Besides frequent algorithmic improvements, each technology generation alters circuit physics and requires new computer-aided design software. The cost of design has doubled in a few years, becoming prohibitive for integrated circuits with limited market penetration[14]. Emerging technologies, such as FinFETs and high-κ dielectrics (κ is the dielectric constant), circumvent known obstacles using forms of design optimization. Therefore, reasonably tight limits should account for potential future optimizations. Low-level technology enhancements, no matter how powerful, are often viewed as one-off improvements, in contrast to architectural redesigns that affect many processor generations. Between technology enhancements and architectural redesigns are global and local optimizations that alter the 'texture' of integrated circuit design, such as logic restructuring, gate sizing and device parameter selection. Moore's law promises higher transistor densities, but some transistors are designed to be 32 times larger than others. Large gates consume greater power to drive long interconnects at acceptable speed and satisfy performance constraints. Minimizing circuit area and power, subject to timing constraints (by configuring each logic gate to a certain size, threshold voltage, and so on), is a difficult but increasingly important optimization with a large parameter space. A recent convex optimization method[28] saved 30% power in Intel chips, and the impact of such improvements grows with circuit size. Many aspects of integrated circuit design are being improved, continually raising the bar for technologies that compete with complementary metal-oxide–semiconductors (CMOSs).

Completing new integrated circuit designs, optimizing them and verifying them requires great effort and continuing innovation; for example, the lack of scalable design automation is a limiting factor for analogue integrated circuits[29,30]. In 1999, bottom-up analysis of digital integrated circuit technologies[15,31] outlined design scaling up to self-contained modules with 50,000 standard cells (each cell contains one to three logic gates), but further scaling was limited by long-range interconnect. In 2010, physical separation of modules became less critical, as large-scale placement optimizations, implemented as software tools, assumed greater responsibility for integrated circuit layout and can now intersperse components of nearby modules[32,33]. In a general trend, powerful design automation[34] frees circuit engineers to focus on microarchitecture[33], but increasingly relies on algorithmic optimization. Until recently, this strategy suffered significant losses in performance[35] and power[36] compared to ideal designs, but has now become both successful and indispensable owing to the rapidly increasing complexity of digital and mixed-signal electronic systems. Hardware and software must now be co-designed and co-verified, with software improving at a faster rate. Platform-based design combines high-level design abstractions with the effective re-use of components and functionalities in engineered systems[37]. Customizable domain-specific computing[8] and domain-specific programming languages[9,10] offload specialization to software running on re-usable hardware platforms.

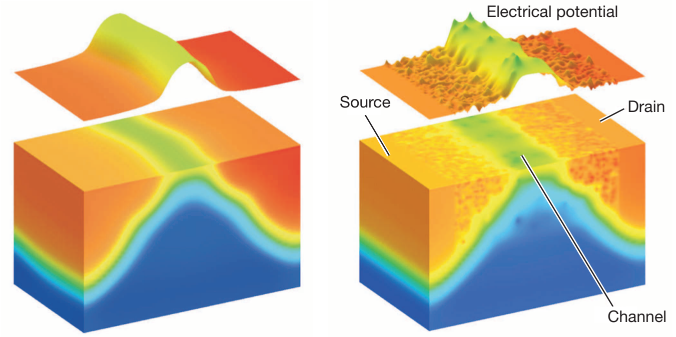


Figure 2 | As a MOSFET transistor shrinks, the shape of its electric field departs from basic rectilinear models, and the level curves become disconnected. Atomic-level manufacturing variations, especially for dopant atoms, start affecting device parameters, making each transistor slightly different[96,97]. Image redrawn from figure 'DOTS and LINES' of ref. 97, with permission from Gold Standard Simulations.

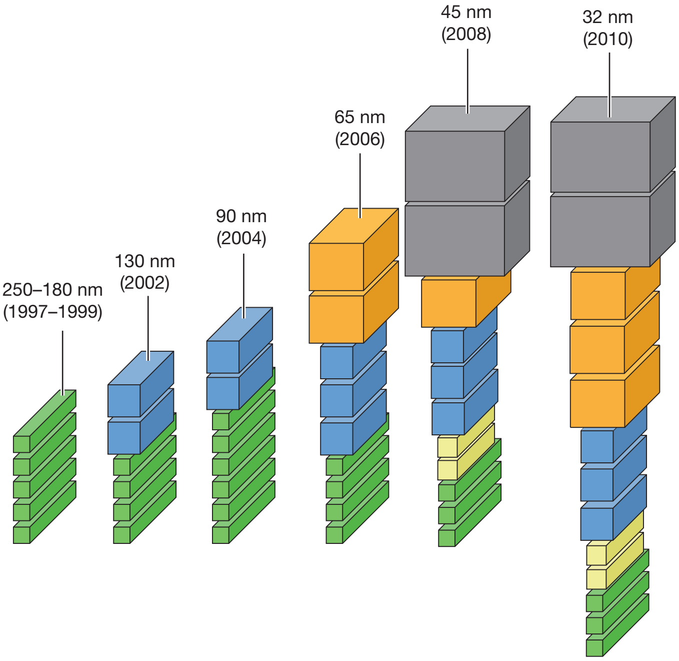


Figure 3 | The evolution of metallic wire stacks from 1997 to 2010. Stacks are ordered by the designation of the semiconductor technology node. Image redrawn from a presentation image by C. Alpert of IBM Research, with permission.

Energy–time limits

In predicting the main obstacles to improving modern electronics, the 2013 edition of the International Technology Roadmap for Semiconductors (ITRS) highlights the management of system power and energy as the main challenge[14]. The faster the computation, the more energy it consumes, but actual power–performance tradeoffs depend on the physical scale. While the ITRS, by its charter, focuses on near-term projections and integrated circuit design techniques, fundamental limits reflect available energy resources, properties of the physical space, power-dissipation constraints, and energy waste.

Reversibility

A 1961 result by Landauer[38] shows that erasing one bit of information entails an energy loss that $\geq kT\ln 2$ (the thermodynamic threshold), where k is the Boltzmann constant and T is the temperature in Kelvin. This principle was validated empirically in 2012 (ref. 39) and seems to motivate reversible computing[40], where all input information is preserved, incurring additional costs. Formally speaking, zero-energy computation is prohibited by

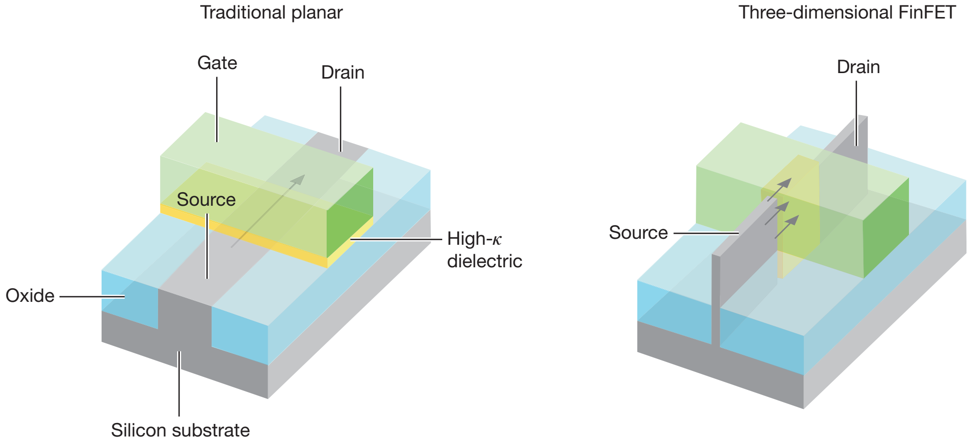


Figure 4 | FinFET transistors possess a much wider gate dielectric layer (surrounding the fin shape) than do MOSFET transistors and can use multiple fins.

the energy–time form of the Heisenberg uncertainty principle ($\Delta t \Delta E \geq \hbar/2$): faster computation requires greater energy[41,42]. However, recent work in applied superconductivity[43] demonstrates "highly exotic" physically reversible circuits operating at 4°K with energy dissipation below the thermodynamic threshold. They apparently fail to scale to large sizes, run into other limits, and remain no more practical than 'mainstream' superconducting circuits and refrigerated low-power CMOS circuits. Technologies that implement quantum circuits[44] can approximate reversible Boolean computing, but currently do not scale to large sizes, are energy-inefficient at the system level, rely on fragile components, and require heavy fault-tolerance overheads[13]. Conventional integrated circuits also do not help to obtain energy savings from reversible computing because they dissipate 30%–60% of all energy in (reversible) wires and repeaters[23]. At room temperature, Landauer's limit amounts to 2.85×10^{-21} J—a very small fraction of the total, given that modern integrated circuits dissipate 0.1–100 W and contain $<10^9$ logic gates. With the increasing dominance of interconnect (see section 'Asymptotic space-time limits'), more energy is spent on communication than on computation. Logically reversible computing is important for reasons other than energy reduction—in cryptography, quantum information processing, and so on[45].

Power constraints and CPUs

The end of CPU frequency scaling. In 2004, Intel abruptly cancelled a 4-GHz CPU project because its high power density required awkward cooling technologies. Other CPU manufacturers kept clock frequencies in the 1–6-GHz range, but also resorted to multicore CPUs[46]. Since dynamic circuit power grows with clock frequency and supply voltage squared[47], energy can be saved by distributing work among slower, lower-voltage parallel CPU cores if the parallelization overhead is small.

Dark, darker, dim, grey silicon. A companion trend to Moore's law—the Dennard scaling theory[48]—shows how to keep power consumption of semiconductor integrated circuits constant while increasing their density. But Dennard scaling broke down ten years ago[48]. Extrapolation of semiconductor scaling trends for CMOSs—the dominant semiconductor technology for the past 20 years—shows that the power consumption of transistors available in modern integrated circuits reduces more slowly than their size (which is subject to Moore's law)[49,50]. To ensure acceptable performance characteristics of transistors, chip power density must be limited, and a fraction of transistors must be kept dark at any given time. Modern CPUs have not been able to use all their circuits at once, but this asymptotic effect—termed the "utilization wall"[49]—will soon black out 99% of the chip, hence the term 'dark silicon' and a reasoned reference to the apocalypse[49]. Saving power by slowing CPU cores down is termed 'dim silicon'. Detailed studies of dark silicon[50] show similar results. To this end, executives from Microsoft and IBM have recently proclaimed an end to the era of multicore microprocessors[51]. Two related trends appeared earlier: (1) increasingly large integrated circuit regions remain transistor-free to aid routeing and physical synthesis, to accommodate power-supply networks, and so on[52,53]—we call them 'darker silicon', (2) increasingly many gates do not perform useful computation but reinforce long, weak interconnects[54] or slow down wires that are too short—which I call 'grey silicon'. Today, 50%–80% of all gates in high-performance integrated circuits are repeaters.

Limits for power supply and cooling. Data centres in the USA consumed 2.2% of its total electricity in 2011. Because power plants take time to build, we cannot sustain past trends of doubled power consumption per year. It is possible to improve the efficiency of transmission lines (using high-temperature superconductors[55]) and power conversion in data centres, but the efficiency of on-chip power networks may soon reach 80%–90%, leaving little room for improvement. Modern integrated circuit power management includes clock-network and power gating[46], per-core voltage scaling[56], charge recovery[57] and, in recent processors, a CPU core dedicated to power scheduling. Integrated circuit power consumption depends quadratically on supply voltage, which has decreased steadily for many years, but has recently stabilized at 0.5–2 V (ref. 47). Supply voltage typically exceeds the threshold voltage of FETs by a safety margin that ensures circuit reliability, fast operation and low leakage. Threshold voltage depends on the thickness of the gate dielectric, which reached a practical limit of several atoms (see section 'Engineering obstacles'). Transistors cannot operate with supply voltage below approximately 200 mV (ref. 17)—five times below current practice—and simple circuits reach this limit. With slower operation, near- and sub-threshold circuits may consume a hundred times less energy[58]. Cooling technologies can improve too, but fundamental quantum limits bound the efficiency of heat removal[59–61].

Broader limits

The study in ref. 62 explores a general binary-logic switch model with binary states represented by two quantum wells separated by a potential barrier. Representing information by electric charge requires energy for binary switching and thus limits the logic-switching density, if a significant fraction of the chip can switch simultaneously. To circumvent this limit, one can encode information in spin-states, photon polarizations, super-conducting currents, or magnetic flux, noting that these carriers have already been in commercial use (spin-states are particularly attractive because they promise high-density nonvolatile storage[63]). More powerful limits are based on the amount of material in the Earth's crust (where silicon is the second most common element after oxygen), on atomic spacing (see section 'Engineering obstacles'), radii, energies and bandgaps, as well as the wavelength of the electron. We are currently using only a tiny fraction of the Earth's mass for computing, and yet various limits could be circumvented if new particles are discovered. Beyond atomic physics, some limits rely on basic constants: the speed of light, the gravitational constant, the quantum (Planck) scale, the Boltzmann constant, and so on. Lloyd[42] and Kraus[64] extend well-known bounds by Bremermann and Bekenstein, and give Moore's law another 150 years and 600 years, respectively. These results are too loose to obstruct the performance of practical computers. In contrast, current consensus estimates from the ITRS[14] give Moore's law only another 10–20 years, due to technological and economic considerations[2].

Asymptotic space–time limits

Engineering limits for deployed technologies can often be circumvented, while first-principles limits on energy and power are loose. Reasonably tight limits are rare.

Limits to parallelism

Suppose we wish to compare a parallel and sequential computer built from the same units, to argue that a new parallel algorithm is many times faster than the best sequential algorithm (the same reasoning applies to logic gates on an integrated circuit). Given N parallel units and an algorithm that runs M times faster on sufficiently large inputs, one can simulate the parallel system on the sequential system by dividing its time between N computational slices. Since this simulation is roughly N times slower, it runs M/N times faster than the original sequential algorithm. If this original sequential algorithm was the fastest possible, we have $M \leq N$. In other words, a fair comparison should not demonstrate a parallel speedup that exceeds the number of processors—a superlinear speedup can indicate an inferior sequential algorithm or the availability of a larger amount of memory to N processors. The bound is reasonably tight in practice for small N and can be violated slightly because N CPUs include more CPU cache, but such violations alone do not justify parallel algorithms—one could instead buy or build one CPU with a larger cache. A linear speedup is optimistically assumed for the parallelizable component in the 1988 Gustafson's law that suggests scaling the number of processors with input size (as illustrated by instantaneous search queries over massive data sets)[5]. Also in 1988, Fisher[65] employed asymptotic runtime estimates instead of numerical limits without considering the parallel and sequential runtime components that were assumed in Amdahl's law[66] and Gustafson's law[5]. Asymptotic estimates neglect leading constants and offer a powerful way to capture nonlinear phenomena occurring at large scale.

Fisher[65] assumes a sequential computation with $T(n)$ elementary steps for input of size n, and limits the performance of its parallel variants that can use an unbounded d-dimensional grid of finite-size computing units (electrical switches on a semiconductor chip, logic gates, CPU cores, and so on) communicating at a finite speed, say, bounded by the speed of light. I highlight only one aspect of this four-page work: the number of steps required by parallel computation grows as the $(d+1)$th root of $T(n)$. This result undermines the N-fold speedup assumed in Gustafson's law for N processors on appropriately sized input data[5]. A speedup from runtime polynomial in n to approximately $\log n$ can be achieved in an abstract model of computation for matrix multiplication and fast Fourier transforms. But not in physical space[65]. Surprising as it may seem, after reviewing many loose limits to computation, we have identified a reasonably tight limit (the impact of input–output, which is a major bottleneck today, is also covered in ref. 65). Indeed, many parallel computations today (excluding multimedia processing and World Wide Web searching) are limited by several forms of communication and synchronization, including network and storage access. The billions of logic gates and memory elements in modern integrated circuits are linked by up to 16 levels of wires (Fig. 3); longer wires are segmented by repeaters. Most of the physical volume and circuit delay are attributed to interconnect[23]. This is relatively new, because gate delays were dominant until 2000 (ref. 14), but wires get slower relative to gates at each new technology node. This uneven scaling has compounded in ways that would have surprised Turing and von Neumann—a single clock cycle is now far too short for a signal to cross the entire chip, and even the distance covered in 200 ps (5 GHz) at light speed is close to the chip size. Yet most electrical engineers and computer scientists are still primarily concerned with gates.

Implications for three-dimensional and other emerging circuits

The promise of three-dimensional integration for improving circuit performance can be undermined by the technical obstructions to its industry adoption. To derive limits on possible improvement, we use the result from ref. 65, which is sensitive to the dimension of the physical space: a sequential computation with $T(n)$ steps requires of the order of $T^{1/3}(n)$ steps in two dimensions and $T^{1/4}(n)$ in three. Letting $t = T^{1/3}(n)$ shows that three-dimensional integration asymptotically reduces t to $t^{3/4}$—a significant but not dramatic speedup. This speedup requires an unbounded number of two-dimensional device layers, otherwise there is no asymptotic speedup[67]. For three-dimensional integrated circuits with two to three layers, the main benefits of three-dimensional integrated circuit integration today are in improving manufacturing yield, improving input–output bandwidth, and combining two-dimensional integrated circuits that are optimized for random logic, dense memory, field-programmable gate-arrays, analogue, microelectromechanical systems and so on. Ultrahigh-density CMOS logic integrated circuits with monolithic three-dimensional integration[68] suffer higher routeing congestion than traditional two-dimensional integrated circuits.

Emerging technologies promise to improve device parameters, but often remain limited by scale, faults, and interconnect. For example, quantum dots enable terahertz switching but hamper nonlocal communication[69]. Carbon nanotube FETs[70] leverage the extraordinary carrier mobility in semiconducting carbon nanotubes to use interconnect more efficiently by improving drive strength, while reducing supply voltage. Emerging interconnects include silicon photonics, demonstrated by Intel in 2013 (ref. 71) and intended as a 100-Gb s^{-1} replacement of copper cables connecting adjacent chips. Silicon photonics promises to reduce power consumption and form factor.

In a different twist, quantum physics alters the nature of communication with Einstein's "spooky action at a distance" facilitated by entanglement[13]. However, the flows of information and entropy are subject to quantum limits[59,60]. Several quantum algorithms run asymptotically faster than the best conventional algorithms[13], but fault-tolerance overhead offsets their potential benefits in practice except for large input sizes, and the empirical evidence of quantum speedups has not been compelling so far[72,73]. Several stages in the development of quantum information processing remain challenging[99], and the surprising difficulty of scaling up reliable quantum computation could stem from limits on communication and entropy[13,59,60]. In contrast, Lloyd[42] notes that individual quantum devices now approach the energy limits for switching, whereas non-quantum devices remain orders of magnitude away. This suggests a possible obstacle to simulating quantum physics on conventional parallel computers (abstract models aside). In terms of computational complexity though, quantum computers cannot attain a significant advantage for many problem types[11–13] and are unlikely to overcome the Fisher limit on parallelism from ref. 65. A similar lack of a consistent general-purpose speedup limits the benefits of several emerging technologies in mature applications that contain diverse algorithmic steps, such as World Wide Web searching and computer-aided design. Accelerating one step usually does not dramatically speed up the entire application, as noted by Amdahl[66] in 1967. Figuratively speaking, the most successful computers are designed for the decathlon rather than for the sprint only.

Complexity-theoretic limits

The previous section, 'Asymptotic space-time limits', enabled tighter limits by neglecting energy and using asymptotic rather than numeric bounds. I now review a more abstract model in order to focus on the impact of scale, and to show how recurring trends quickly overtake one-off device-specific effects. I neglect spatial effects and focus on the nature of computation in an abstract model (used by software engineers) that represents computation by elementary steps with input-independent runtimes. Such limits survive many improvements in computer technologies, and are often stronger for specific problems. For example, the best-known algorithms for multiplying large numbers are only slightly slower than reading the input (an obvious speed limit), but only in the asymptotic sense: for numbers with less than a thousand bits, those algorithms lag behind simpler algorithms in actual performance. To focus on what matters most, I no longer track the asymptotic worst-case complexity of the best algorithms for a given problem, but merely distinguish polynomial asymptotic growth from exponential.

Limits formulated in such crude terms (unsolvability in polynomial time on any computer) are powerful[74]: the hardness of number-factoring underpins Internet commerce, while the P ≠ NP conjecture explains the lack of satisfactory, scalable solutions to important algorithmic problems, in optimization and verification of integrated circuit designs, for example[75]. (Here P is the class of decision problems that can be solved using simple computational steps whose number grows no faster than a polynomial of the size of input data, and NP is the non-deterministic polynomial class representing those decision problems for which a non-deterministically guessed solution can be reliably checked using a polynomial number of steps.) A similar conjecture, P ≠ NC, seeks to explain why many algorithmic problems that can be solved efficiently have not parallelized efficiently[76]. Most of these limits have not been proved. Some can be circumvented by using radically different physics, for example, quantum computers can solve number factoring in polynomial time (in theory). But quantum computation does not affect P ≠ NP (ref. 77). The lack of proofs, despite heavy empirical evidence, requires faith and is an important limitation of many nonphysical limits to computing. This faith is not universally shared—Knuth (see question 17 in http://www.informit.com/articles/article.aspx?p=2213858) argues that P = NP would not contradict anything we know today. A rare proved result by Turing states that checking whether a given program ever halts is undecidable: no algorithm solves this problem in all cases regardless of runtime. Yet software developers solve this problem during peer code reviews, and so do computer science teachers when grading exams in programming courses.

Worst-case analysis is another limitation of nonphysical limits to computing, but suggests potential gains through approximation and specialization. For some NP-hard optimization problems, such as the Euclidean Travelling Salesman Problem, polynomial-time approximations exist, but in other cases, such as the Maximum Clique problem, accurate approximation is as hard as finding optimal solutions[78]. For some important problems and algorithms, such as the Simplex algorithm for linear programming, few inputs lead to exponential runtime, and minute perturbations reduce runtime to polynomial[79].

Conclusions

The death march of Moore's law[1,2] invites discussions of fundamental limits and alternatives to silicon semiconductors[70]. Near-term constraints (obstacles to performance, power, materials, laser sources, manufacturing technologies and so on) are invariably tied to costs and capital, but are disregarded for the moment as new markets for electronics open up, populations increase, and the world economy grows[2]. Such economic pressures emphasize the value of computational universality and the broad applicability of integrated circuit architectures to solve multiple tasks under conventional environmental conditions. In a likely scenario, only CPUs, graphics processing units, field-programmable gate-arrays and dense memory integrated circuits will remain viable at the end of Moore's law, while specialized circuits will be predominantly manufactured with less advanced technologies for financial reasons. Indeed, memory chips have exemplified Moore scaling because of their simpler structure, modest interconnect, and more controllable manufacturing, but the miniaturization of memory cells is now slowing down[2]. The decelerated scaling of CMOS integrated circuits still outperforms the scaling of the most viable emerging technologies. Empirical scaling laws describing the evolution of computing are well known[80]. In addition to Moore's law, Dennard scaling, Amdahl's law and Gustafson's law (reviewed above), Metcalfe's law[81] states that the value of a computer network, such as the Internet or Facebook, scales as the number of user-to-user connections that can be formed. Grosch's law[82] ties N-fold improvements in computer performance to N^2-fold cost increases (in equivalent units). Applying it in reverse, we can estimate the acceptable performance of cheaper computers. However, such laws only capture ongoing scaling and may not apply in the future.

The roadmapping process represented by the ITRS[14] relies on consensus estimates and works around engineering obstacles. It tracks improvements in materials and tools, collects best practices and outlines promising design strategies. As suggested in refs 17 and 18, it can be enriched by an analysis of limits. I additionally focus on how closely such limits can be approached. Aside from the historical 'wrong turns' mentioned in the 'Engineering obstacles' and 'Energy–time limits' sections above, I uncover interesting effects when examining the tightness of individual limits. Although energy–time limits are most critical in computer design[14,83], space-time limits appear tighter[65] and capture bottlenecks formed by interconnect and communication. They suggest optimizing gate locations and sizes, and placing gates in three dimensions. One can also adapt algorithms to spatial embeddings[84,85] and seek space-time limits. But the gap between current technologies and energy–time limits hints at greater possible rewards. Charge recovery[57], power management[46], voltage scaling[56], and near-threshold computing[58] reduce energy waste. Optimizing algorithms and circuits simultaneously for energy and spatial embedding[86] gives biological systems an edge (from the 'one-dimensional' nematode *Caenorhabditis elegans* with 302 neurons to the three-dimensional human brain with 86 billion neurons)[1]. Yet, using the energy associated with mass (according to Einstein's $E = mc^2$ formula) to compute can truly be a 'nuclear option'—both powerful and controversial. In a well known 1959 talk, which predated Moore's law, Richard Feynman suggested that there was "plenty of room at the bottom," forecasting the miniaturization of electronics. Today, with relatively little physical room left, there is plenty of energy at the bottom. If this energy is tapped for computing, how can the resulting heat be removed? Recycling heat into mass or electricity seems to be ruled out by limits to energy conversion and the acceptable thermal range for modern computers.

Technology-specific limits for modern computers tend to express trade-offs, especially for systems with conflicting performance parameters and properties[87]. Little is known about limits on design technologies. Given that large-scale complex systems are often designed and implemented hierarchically[52] with multiple levels of abstraction, it would be valuable to capture losses incurred at abstraction boundaries (for example, the physical layout and manufacturing considerations required to optimize and build a logic circuit may mean that the logic circuit itself needs to change) and between levels of design hierarchies. It is common to estimate resources required for a subsystem and then to implement the subsystem to satisfy resource budgets. Underestimation is avoided because it leads to failures, but overestimation results in overdesign. Inaccuracies in estimation and physical modelling also lead to losses during optimization, especially in the presence of uncertainty. Clarifying engineering limits gives us the hope of circumventing them.

Technology-agnostic limits appear to be simple and have had significant effects in practice; for example, Aaronson explains why NP-hardness is unlikely to be circumvented through physics[77]. Limits to parallel computation became prominent after CPU speed levelled off ten years ago. These limits suggest that it will be helpful to use the following: faster interconnect[18], local computation that reduces communication[88], time-division multiplexing of logic[89], architectural and algorithmic techniques[90], and applications altered to embrace parallelism[5]. Gustafson advocates a 'natural selection': the survival of the applications that are fittest for parallelism. In another twist, the performance and power consumption of industry-scale distributed systems is often described by probability distributions, rather than single numbers[91,92], making it harder even to formulate appropriate limits. We also cannot yet formulate fundamental limits related to the complexity of the software-development effort, the efficiency of CPU caches[93], and the computational requirements of incremental functional verification, but we have noticed that many known limits are either loose or can be circumvented, leading to secondary limits. For example, the P ≠ NP limit is worded in terms of worst-case rather than average-case performance, and has not been proved despite much empirical evidence. Researchers have ruled out entire categories of proof techniques as insufficient to complete such a proof[75,94]. They may be esoteric, but such tertiary limits can be effective in practice—in August 2010, they helped researchers quickly invalidate Vinay Deolalikar's highly technical attempt at proving

P ≠ NP. On the other hand, the correctness of lengthy proofs for some key results could not be established with an acceptable level of certainty by reviewers, prompting efforts towards verifying mathematics by computation[95].

In summary, I have reviewed what is known about limits to computation, including existential challenges arising in the sciences, optimization challenges arising in engineering, and the current state of the art. These categories are closely linked during rapid technology development. When a specific limit is approached and obstructs progress, understanding its assumptions is a key to circumventing it. Some limits are hopelessly loose and can be ignored, while other limits remain conjectural and are based on empirical evidence only; these may be very difficult to establish rigorously. Such limits on limits to computation deserve further study.

Received 21 December 2013; accepted 5 June 2014.

1. Cavin, R. K., Lugli, P. & Zhirnov, V. V. Science and engineering beyond Moore's law. *Proc. IEEE* **100,** 1720–1749 (2012).
This paper reviews the historical effects and benefits of Moore's law, discusses challenges to further growth, and offers several strategies to maintain progress in electronics.
2. Chien, A. A. & Karamcheti, V. Moore's law: the first ending and a new beginning. *IEEE Computer* **46,** 48–53 (2013).
3. Herken, R. (ed.) *The Universal Turing Machine: A Half-Century Survey* 2nd edn (Springer, 2013).
4. Andreesen, M. Why software is eating the world. *The Wall Street Journal* http://online.wsj.com/news/articles/SB10001424053111903480904576512250915629460(11 (August 2011).
5. Padua, D. A. (ed.) *Encyclopedia of Parallel Computing* (Springer, 2011).
6. Shaw, D. E. Anton: a special-purpose machine that achieves a hundred-fold speedup in biomolecular simulations. In *Proc. Int. Symp. on High Performance Distributed Computing* 129–130 (IEEE, 2013).
7. Hameed, R. *et al.* Understanding sources of inefficiency in general-purpose chips. *Commun. ACM* **54,** 85–93 (2011).
8. Cong, J., Reinman, G., Bui, A. T. & Sarkar, V. Customizable domain-specific computing. *IEEE Des. Test Comput.* **28,** 6–15 (2011).
9. Mernik, M., Heering, J. & Sloane, A. M. When and how to develop domain-specific languages. *ACM Comput. Surv.* **37,** 316–344 (2005).
10. Olukotun, K. Beyond parallel programming with domain specific languages. In *Proc. Symp. on Principles and Practice of Parallel Programming* 179180 (ACM, 2014).
11. Aaronson, S. & Shi, Y. Quantum lower bounds for the collision and the element distinctness problems. *J. ACM* **51,** 595–605 (2004).
12. Jain, R., Ji, Z., Upadhyay, S. & Watrous, J. QIP = PSPACE. *Commun. ACM* **53,** 102–109 (2010).
13. Nielsen, M. A. & Chuang, I. L. *Quantum Computation and Quantum Information* (Cambridge Univ. Press, 2011).
14. International Technology Roadmap for Semiconductors (ITRS). http://www.itrs.net/ (2013).
Documents available at this website describe in detail the current state of the art in integrated circuits and near-term milestones.
15. Sylvester, D. & Keutzer, K. A global wiring paradigm for deep submicron design. *IEEE Trans. CAD* **19,** 242–252 (2000).
16. *A Quantum Information Science and Technology Roadmap* Los Alamos Technical Report LA-UR-04–1778, http://qist.lanl.gov (2004).
17. Meindl, J. Low power microelectronics: retrospective and prospect. *Proc. IEEE* **83,** 619–635 (1995).
18. Davis, J. A. *et al.* Interconnect limits on gigascale integration (GSI) in the 21st Century. *Proc. IEEE* **89,** 305–324 (2001).
This paper discusses physical limits to scaling interconnects in integrated circuits, classifying them into fundamental, material, device, circuit and system limits.
19. Ma, X. & Arce, G. R. *Computational Lithography* (Wiley, 2011).
20. Mazzola, L. Commercializing nanotechnology. *Nature Biotechnol.* **21,** 1137–1143 (2003).
21. Moore, G. E. Cramming more components onto integrated circuits. *Electronics* **38,** 1–4 (1965).
22. Bohr, M. Interconnect scaling—the real limiter to high performance ULSI. In *Proc. Int. Elec. Device Meeting* 241–244 (IEEE, 1995).
This paper explains why wires, rather than gates, have become the main limiter to the performance of ultra-large integrated circuits.
23. Shelar, R. & Patyra, M. Impact of local interconnects on timing and power in a high performance microprocessor. *IEEE Trans. CAD* **32,** 1623–1627 (2013).
24. Almeida, V. R., Barrios, C. A., Panepucci, R. R. & Lipson, M. All-optical control of light on a silicon chip. *Nature* **431,** 1081–1084 (2004).
25. Chang, M.-C. F., Roychowdhury, V. P., Zhang, L., Shin, H. & Qian, Y. RF/wireless interconnect for inter-and intra-chip communications. *Proc. IEEE* **89,** 455–466 (2002).
26. Hisamoto, D. *et al.* FinFET—a self-aligned double-gate MOSFET scalable to 20 nm. *IEEE Trans. Electron. Dev.* **47,** 2320–2325 (2002).
27. Seabaugh, A. The tunneling transistor. *IEEE Spectrum* http://spectrum.ieee.org/semiconductors/devices/the-tunneling-transistor (2013).
28. Ozdal, M. M., Burns, S. M. & Hu, J. Algorithms for gate sizing and device parameter selection for high-performance designs. *IEEE Trans. CAD* **31,** 1558–1571 (2012).
29. Rutenbar, R. A. Design automation for analog: the next generation of tool challenges. In *Proc. Int. Conf. Computer-Aided Design of Integrated Circuits* 458–460 (IEEE, 2006).
30. Rutenbar, R. A. Analog layout synthesis: what's missing? In *Proc. Int. Symp. Physical Design of Integrated Circuits* 43 (ACM, 2010).
31. Ho, R., Mai, K., Kapadia, H. & Horowitz, M. Interconnect scaling implications for CAD. In *Proc. Int. Conf. Computer-Aided Design of Integrated Circuits* 425–429 (IEEE, 1999).
32. Markov, I. L., Hu, J. & Kim, M.-C. Progress and challenges in VLSI placement research. In *Proc. Int. Conf. Computer-Aided Design of Integrated Circuits* 275–282 (IEEE, 2012).
33. Puri, R. Opportunities and challenges for high-performance CPU designs and design automation. In *Proc. Int. Symp. Physical Design of Integrated Circuits* 179 (ACM, 2013).
34. Lavagno, L., Martin, G. & Scheffer, L. *Electronic Design Automation for Integrated Circuits Handbook* (CRC Press, 2006).
35. Chinnery, D. G. & Keutzer, K. *Closing the Gap Between ASIC and Custom: Tools And Techniques For High-Performance ASIC Design* (Springer, 2004).
36. Chinnery, D. G. & Keutzer, K. *Closing the Power Gap between ASIC and Custom: Tools and Techniques for Low Power Design* (Springer, 2007).
37. Sangiovanni-Vincentelli, A. L., Carloni, L. P., De Bernardinis, F. & Sgroi, M. Benefits and challenges for platform-based design. In *Proc. Design Automation Conf.* 409–414 (ACM, 2004).
38. Landauer, R. Irreversibility and heat generation in the computing process. *IBM J. Res. Develop.* **5,** 183–191 (1961).
39. Bérut, A. *et al.* Experimental verification of Landauer's principle linking information and thermodynamics. *Nature* **483,** 187–189 (2012).
40. Bennett, C. H. & Landauer, R. The fundamental limits of computation. *Sci. Am.* **253,** 48–56 (1985).
41. Aharonov, Y. & Bohm, D. Time in the quantum theory and the uncertainty relation for time and energy. *Phys. Rev.* **122,** 1649–1658 (1961).
42. Lloyd, S. Ultimate physical limits on computation. *Nature* **406,** 1047–1054 (2000).
This paper derives several *ab initio* limits to computation, points out that modern quantum devices operate close to their energy-efficiency limits, but concludes that resulting large-scale limits are very loose.
43. Ren, J. & Semenov, V. K. Progress with physically and logically reversible superconducting digital circuits. *IEEE Trans. Appl. Supercond.* **21,** 780–786 (2011).
44. Monroe, C. *et al.* Large-scale modular quantum-computer architecture with atomic memory and photonic interconnects. *Phys. Rev. A* **89,** 022317 (2014).
45. Saeedi, M. & Markov, I. L. Synthesis and optimization of reversible circuits — a survey. *ACM Comput. Surv.* **45** (2), 21 (2013).
46. Borkar, S. Thousand-core chips: a technology perspective. In *Proc. Design Automation Conf.* 746–749 (ACM, 2007).
This paper describes Intel's thousand-core CPU architecture with an emphasis on fine-grain power management, memory bandwidth, on-die networks, and system resiliency.
47. Rabaey, J. M., Chandrakasan, A. & Nikolic, B. *Digital Integrated Circuits A Design Perspective* (Pearson Education, 2003).
48. Bohr, M. A 30 year retrospective on Dennard's MOSFET scaling paper. *IEEE Solid-State Circ. Soc. Newsl.* **12,** 11–13 (2007).
This paper reviews power scaling of integrated circuits, which held for 30 years, but has now broken down.
49. Taylor, M. B. Is dark silicon useful? harnessing the four horsemen of the coming dark silicon apocalypse. In *Proc. Design Automation Conf.* 1131–1136 (ACM, 2012).
50. Esmaeilzadeh, H., Blem, E. R., St.-Amant, R., Sankaralingam, K. & Burger, D. Power challenges may end the multicore era. *Commun. ACM* **56,** 93–102 (2013).
51. Yeraswork, Z. 3D stacks and security key for IBM in server market. *EE Times* http://www.eetimes.com/document.asp?doc_id=1320403 (17 December 2013).
52. Caldwell, A. E., Kahng, A. B. & Markov, I. L. Hierarchical whitespace allocation in top-down placement. *IEEE Trans. Computer-Aided Design Integrated Circ.* **22,** 716–724 (2003).
53. Adya, S. N., Markov, I. L. & Villarrubia, P. G. On whitespace and stability in physical synthesis. *Integration VLSI J.* **39,** 340–362 (2006).
54. Saxena, P., Menezes, N., Cocchini, P. & Kirkpatrick, D. Repeater scaling and its impact on CAD. *IEEE Trans. Computer-Aided Design Integrated Circ.* **23,** 451–463 (2004).
55. Oestergaard, J., Okholm, J., Lomholt, K. & Toennesen, G. Energy losses of superconducting power transmission cables in the grid. *IEEE Trans. Appl. Supercond.* **11,** 2375 (2001).
56. Pinckney, N. R. *et al.* Limits of parallelism and boosting in dim silicon. *IEEE Micro* **33,** 30–37 (2013).
57. Kim, S., Ziesler, C. H. & Papaefthymiou, M. C. Charge-recovery computing on silicon. *IEEE Trans. Comput.* **54,** 651–659 (2005).
58. Dreslinski, R. G., Wieckowski, M., Blaauw, D., Sylvester, D. & Mudge, T. Near-threshold computing: reclaiming Moore's law through energy efficient integrated circuits. *Proc. IEEE* **98,** 253–266 (2010).
59. Pendry, J. B. Quantum limits to the flow of information and entropy. *J. Phys. Math. Gen.* **16,** 2161–2171 (1983).
60. Blencowe, M. P. & Vitelli, V. Universal quantum limits on single-channel information, entropy, and heat flow. *Phys. Rev. A* **62,** 052104 (2000).
61. Whitney, R. S. Most efficient quantum thermoelectric at finite power output. *Phys. Rev. Lett.* **112,** 130601 (2014).
62. Zhirnov, V. V., Cavin, R. K., Hutchby, J. A. & Bourianoff, G. I. Limits to binary logic switch scaling—a Gedanken model. *Proc. IEEE* **91,** 1934–1939 (2003).

63. Wolf, S. A. *et al.* Spintronics: a spin-based electronics vision for the future. *Science* **294,** 1488–1495 (2001).
64. Krauss, L. M. & Starkman, G. D. Universal limits on computation. Preprint at http://arxiv.org/abs/astro-ph/0404510 (2004).
65. Fisher, D. Your favorite parallel algorithms might not be as fast as you think. *IEEE Trans. Comput.* **37,** 211–213 (1988).
66. Amdahl, G. M. Computer architecture and Amdahl's law. *IEEE Computer* **46,** 38–46 (2013).
67. Mak, W.-K. & Chu, C. Rethinking the wirelength benefit of 3-D integration. *IEEE Trans. VLSI Syst.* **20,** 2346–2351 (2012).
68. Lee, Y.-J., Morrow, P. & Lim, S. K. Ultra high density logic designs using transistor-level monolithic 3D integration. In *Proc. Int. Conf. Computer-Aided Design of Integrated Circuits* 539–546 (IEEE, 2012).
69. Sherwin, M. S., Imamoglu, A. & Montroy, Th. Quantum computation with quantum dots and terahertz cavity quantum electrodynamics. *Phys. Rev. A* **60,** 3508 (1999).
70. Shulaker, M. *et al.* Carbon nanotube computer. *Nature* **501,** 526–530 (2013).
71. Simonite, T. Intel's laser chips could make data centers run better. *MIT Technol. Rev.* (4 September 2013).
72. Rønnow, T. F. *et al.* Defining and detecting quantum speedup. *Science* **20,** 1330–1331 (2014).
This paper shows how to define and measure quantum speedup, while avoiding pitfalls and overly optimistic results—an empirical study with a D-Wave 2 chip with up to 503 qubits finds no convincing evidence of speed-up.
73. Shin, S. W., Smith, G., Smolin, J. A. & Vazirani, U. How 'quantum' is the D-Wave machine? Preprint at http://arxiv.org/abs/1401.7087 (2014).
74. Sipser, M. *Introduction to the Theory of Computation* 3rd edn (Cengage Learning, 2012).
75. Fortnow, L. The status of the P versus NP problem. *Commun. ACM* **52,** 78–86 (2009).
76. Markov, I. L. Know your limits: a review of 'limits to parallel computation: P-completeness theory'. *IEEE Design Test* **30,** 78–83 (2013).
77. Aaronson, S. Guest column: NP-complete problems and physical reality. *SIGACT (ACM Special Interest Group on Algorithms and Computation Theory) News* **36,** 30–52 (2005).
This paper explores the possibility of efficiently solving NP-complete problems using analog, adiabatic and quantum computing, protein folding and soap bubbles, as well as other proposals for physics-based computing— it concludes that this is unlikely, but suggests other benefits of studying such proposals.
78. Vazirani, V. *Approximation Algorithms* (Springer, 2002).
79. Spielman, D. & Teng, S.-H. Smoothed analysis of algorithms: why the simplex algorithm usually takes polynomial time. In *Proc. Symp. Theory of Computing* 296305 (ACM, 2001).
80. Getov, V. Computing laws: origins, standing, and impact. *IEEE Computer* **46,** 24–25 (2013).
81. Metcalfe, B. Metcalfe's law after 40 years of ethernet. *IEEE Computer* **46,** 26–31 (2013).
82. Ryan, P. S., Falvey, S. & Merchant, R. When the cloud goes local: the global problem with data localization. *IEEE Computer* **46,** 54–59 (2013).
83. Wenisch, T. F. & Buyuktosunoglu, A. Energy-aware computing. *IEEE Micro* **32,** 6–8 (2012).
84. Bachrach, J. & Beal, J. Developing spatial computers. Technical Report MITCSAIL-TR-2007–017 (MIT, 2007).
85. Rosenbaum, D. Optimal quantum circuits for nearest-neighbor architectures. Preprint at http://arxiv.org/abs/1205.0036; in *8th Conf. on the Theory of Quantum Computation, Communication and Cryptography* 294–307 (Schloss Dagstuhl—Leibniz-Zentrum fuer Informatik, 2012).
86. Patil, D., Azizi, O., Horowitz, M., Ho, R. & Ananthraman, R. Robust energy-efficient adder topologies. In *Proc. IEEE Symp. on Computer Arithmetic* 16–28 (IEEE, 2007).
87. Brewer, E. CAP twelve years later: how the 'rules' have changed. *IEEE Computer* **45,** 23–29 (2012).
88. Demmel, J. Communication-avoiding algorithms for linear algebra and beyond. In *Proc. Int. Parallel and Distributed Processing Symp.* 585 (IEEE, 2013).
89. Halfill, T. R. Tabula's time machine. *Microprocessor Report* (29 March 2010).
90. Dror, R. O. *et al.* Overcoming communication latency barriers in massively parallel scientific computation. *IEEE Micro* **31,** 8–19 (2011).
91. Dean, J. & Barroso, L. A. The tail at scale. *Commun. ACM* **56,** 74–80 (2013).
92. Barroso, L. A., Clidaras, J. & Hölzle, U. *The Datacenter as a Computer: An Introduction to the Design of Warehouse-Scale Machines* 2nd edn (Synthesis Lectures on Computer Architecture, Morgan & Claypool, 2013).
93. Balasubramonian, R., Jouppi, N. P. & Muralimanohar, N. *Multi-Core Cache Hierarchies* (Synthesis Lectures on Computer Architecture, Morgan & Claypool, 2011).
94. Aaronson, S. & Wigderson, A. Algebrization: a new barrier in complexity theory. *ACM Trans. Complexity Theory* **1** (1), http://www.scottaaronson.com/papers/alg.pdf (2009).
95. Avigad, J. & Harrison, J. Formally verified mathematics. *Commun. ACM* **57,** 66–75 (2014).
96. Asenov, A. Random dopant induced threshold voltage lowering and fluctuations in sub-0.1 m MOSFETs: a 3-D ''atomistic'' simulation study. *IEEE Trans. Electron. Dev.* **45,** 2505–2513 (1998).
97. Miranda, M. The threat of semiconductor variability. *IEEE Spectrum* http://spectrum.ieee.org/semiconductors/design/the-threat-of-semiconductor-variability (2012).
98. Naeemi, A. *et al.* BEOL scaling limits and next generation technology prospects. *Proc. Design Automation Conf.* 1–6 (ACM, 2014).
99. Devoret, M. H. & Schoelkopf, R. J. Superconducting circuits for quantum information: an outlook. *Science* **339,** 1169–1173 (2013).

Acknowledgements This work was supported in part by the Semiconductor Research Corporation (SRC) Task 2264.001 (funded by Intel and IBM), a US Airforce Research Laboratory award (FA8750-11-2-0043), and a US National Science Foundation (NSF) award (1162087).

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to the author (imarkov@eecs.umich.edu).

doi:10.1038/nature13600

Clonal evolution in breast cancer revealed by single nucleus genome sequencing

Yong Wang[1], Jill Waters[1], Marco L. Leung[1,2], Anna Unruh[1], Whijae Roh[1], Xiuqing Shi[1], Ken Chen[3], Paul Scheet[2,4], Selina Vattathil[2,4], Han Liang[3], Asha Multani[1], Hong Zhang[5], Rui Zhao[6], Franziska Michor[6], Funda Meric-Bernstam[7] & Nicholas E. Navin[1,2,3]

Sequencing studies of breast tumour cohorts have identified many prevalent mutations, but provide limited insight into the genomic diversity within tumours. Here we developed a whole-genome and exome single cell sequencing approach called nuc-seq that uses G2/M nuclei to achieve 91% mean coverage breadth. We applied this method to sequence single normal and tumour nuclei from an oestrogen-receptor-positive (ER^+) breast cancer and a triple-negative ductal carcinoma. In parallel, we performed single nuclei copy number profiling. Our data show that aneuploid rearrangements occurred early in tumour evolution and remained highly stable as the tumour masses clonally expanded. In contrast, point mutations evolved gradually, generating extensive clonal diversity. Using targeted single-molecule sequencing, many of the diverse mutations were shown to occur at low frequencies (<10%) in the tumour mass. Using mathematical modelling we found that the triple-negative tumour cells had an increased mutation rate (13.3×), whereas the ER^+ tumour cells did not. These findings have important implications for the diagnosis, therapeutic treatment and evolution of chemoresistance in breast cancer.

Human breast cancers often display intratumour genomic heterogeneity[1–3]. This clonal diversity confounds the clinical diagnosis and basic research of human cancers. Expression profiling has shown that breast cancers can be classified into five molecular subtypes that correlate with the presence of oestrogen, progesterone and Her2 receptors[4]. Among these, triple-negative breast cancers ($ER^-/PR^-/Her2^-$) have been shown to harbour the largest number of mutations, whereas luminal A ($ER^+/PR^+/Her2^-$) breast cancers show the lowest frequencies[5–7]. These data suggest that triple-negative breast cancers (TNBCs) may have increased clonal diversity and mutational evolution, but such inferences are difficult to make in bulk tissues[8,9]. To gain better insight into the genomic diversity of breast tumours, we developed a single cell genome sequencing method and applied it to study mutational evolution in an ER^+ breast cancer (ERBC) and a TNBC patient. We combined this approach with targeted duplex[10] single-molecule sequencing to profile thousands of cells and understand the role of rare mutations in tumour evolution.

Whole-genome sequencing using G2/M nuclei

In our previous work we developed a method using degenerate-oligonucleotide PCR and sparse sequencing to measure copy number profiles of single cells[11]. Although adequate for copy number detection, this method could not resolve genome-wide mutations at base-pair resolution. We attempted to increase coverage by deep-sequencing these libraries, but found that coverage breadth approached a limit near 10% (Fig. 1a). To address this problem, we developed a high-coverage, whole-genome and exome single cell sequencing method called nuc-seq (Extended Data Fig. 1). In this method we exploit the natural cell cycle, in which single cells duplicate their genome during S phase, expanding their DNA from 6 to 12 picograms before cytokinesis. This approach provides an advantage over using chemical inhibitors to induce polyploidy in single cells[12,13] because it does not require live cells.

We input four (or more) copies of each single cell genome for whole-genome-amplification (WGA) to decrease the allelic dropout and false positive error rates, which are major sources of error during multiple-displacement amplification (MDA)[14,15]. Additionally, we limit the MDA time to 80 min to mitigate false positive (FP) errors associated with the infidelity of the ϕ29 polymerase (Methods). The improved amplification efficiency can be shown using 22 chromosome-specific primer pairs for PCR (Extended Data Fig. 2). In G1/G0 single cells we find that only 25.58% (11/43) of the cells show full amplification of the chromosomes, whereas G2/M cells have 45.34% (39/86). After MDA, we incubate the amplified DNA with a Tn5 transposase, which simultaneously fragments DNA and ligates adapters for sequencing[16]. The libraries are then multiplexed for exome capture or used directly for next-generation sequencing.

Method validation in a monoclonal cancer cell line

To validate our method we used a breast cancer cell line (SK-BR-3) that was previously shown to be genetically monoclonal[11,17]. We evaluated the genetic homogeneity of this cell line using spectral karyotyping and found that large chromosome rearrangements were highly stable in 85.80% of the single cells (Supplementary Table 1). We also performed single nucleus sequencing (SNS)[11,18] on 50 single SK-BR-3 cells and calculated copy number profiles at 220 kilobase (kb) resolution, which showed that the major amplifications of *MET, MYC, ERBB2, BCAS1* and a deletion in *DCC* were stable (mean $R^2 = 0.91$) in all of the 50 cells (Fig. 1b). Next, we deep-sequenced the SK-BR-3 cell population (SKP) at high coverage depth (51×) and breadth (90.40%) and detected single-nucleotide variants (SNVs), copy number aberrations (CNAs) and structural variants (SVs) using our processing pipeline (Methods). We filtered the variants using dbSNP135 and identified 409 non-synonymous variants and 1,452 structural variants (Fig. 1d), several of which occurred in cancer genes (Supplementary Table 2).

We applied nuc-seq to sequence the whole genomes of two single SK-BR-3 cells (SK1 and SK2) and calculated coverage depth, breadth (sites with at least one read) and uniformity (evenness). We found that both SK-BR-3 cells achieved high coverage depth (61× ± 5 s.e.m., $n = 2$)

[1] The University of Texas MD Anderson Cancer Center, Department of Genetics, Houston, Texas 77030, USA. [2] The University of Texas Graduate School of Biomedical Sciences, Houston, Texas 77030, USA. [3] The University of Texas MD Anderson Cancer Center, Department of Bioinformatics and Computational Biology, Houston, Texas 77030, USA. [4] The University of Texas MD Anderson Cancer Center, Department of Epidemiology, Houston, Texas 77030, USA. [5] The University of Texas MD Anderson Cancer Center, Department of Pathology, Houston, Texas 77030, USA. [6] Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute, and Department of Biostatistics, Harvard School of Public Health, Boston, Massachusetts 02215, USA. [7] The University of Texas MD Anderson Cancer Center Department of Investigational Cancer Therapeutics, Houston, Texas 77030, USA.

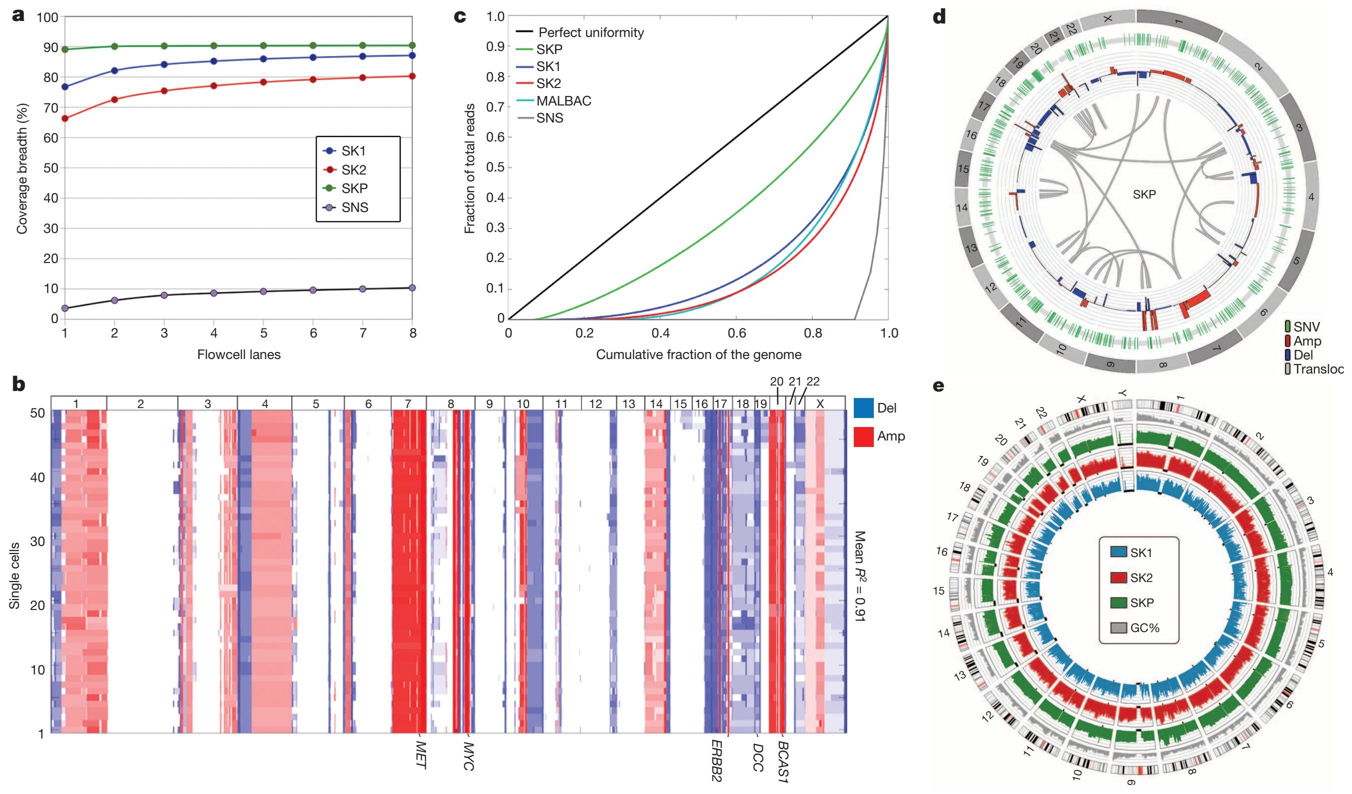


Figure 1 | Method performance in a monoclonal cell line. a, Coverage breadth for single cells (SK1, SK2) sequenced by nuc-seq, a single cell SNS library and a SK-BR-3 population (SKP) sample. **b**, Heatmap of 50 single cell SK-BR-3 copy number profiles. **c**, Lorenz curve of coverage uniformity for the single SK-BR-3 cells sequenced by nuc-seq, a cell sequenced by SNS, a population of SK-BR-3 cells, and a cell sequenced by MALBAC. **d**, Circos plot of variants detected by sequencing populations of SK-BR-3 cells. **e**, Coverage depth for the SK-BR-3 population sample and the SK1 and SK2 single cells.

and breadth (83.70 ± 3.40% s.e.m., $n = 2$) (Fig. 1e). In comparison, we re-analysed coverage breadth in single cells sequenced by MALBAC[19] using unique reads and calculated 69.54% coverage breadth. We evaluated coverage uniformity using Lorenz curves[20] which showed highly uniform coverage, representing a major improvement over our previous SNS method[11,18] and is equivalent to the MALBAC data[19] (Fig. 1c). Next, we calculated error rates, including the allelic dropout rate (ADR) and false positive rate (FPR) by comparing single cell variants to the population data (Methods). Our analysis suggests that nuc-seq generates low allelic dropout rates (9.73 ± 2.19%) compared to previous studies (7–46%)[14]. We also achieved low false positive error rates for point mutations (FPR = 1.24×10^{-6}), equivalent to 1–2 errors per million bases, which represents a major technical improvement over previous methods[14,19] (FPR = 2.52×10^{-5} and 4×10^{-5}).

Population and single nuclei sequencing of an ERBC

We selected an invasive ductal carcinoma from an oestrogen-receptor positive (ER$^+$/PR$^+$/Her2$^-$) breast cancer patient for population and single cell sequencing (Fig. 2a, Methods). We flow-sorted millions of nuclei from the aneuploid G2/M peak (6N) and from matched normal tissue for population sequencing (46× and 54×) (Fig. 2b). We also flow-sorted 50 single nuclei for copy number profiling, 4 nuclei for whole-genome sequencing and 59 nuclei for exome sequencing. After filtering germline variants, we identified a total of 4,162 somatic SNVs in the aneuploid tumour cell population. Among these SNVs we identified 12 non-synonymous mutations, which we validated by exome sequencing (66×). Several non-synonymous mutations occurred in cancer genes, including *PIK3CA*, *CASP3*, *FBN2* and *PPP2R5E* (Fig. 2c, Supplementary Information). *PIK3CA* is the most common driver mutation in luminal A breast cancers[7,9].

To investigate copy number diversity, we performed single nucleus sequencing[11,18] on 50 single nuclei. We constructed a neighbour-joining tree, which showed that single tumour cells shared highly similar CNAs (mean $R^2 = 0.89$), representing a monoclonal population (Fig. 2d, Extended Data Fig. 3a). Next, we performed whole-genome sequencing of four single tumour nuclei at high coverage breadth (80.79 ± 3.31% s.e.m., $n = 4$) and depth (mean 46.75× ± 5.06 s.e.m., $n = 4$). From this data we identified three classes of mutations: (1) clonal mutations, detected in the population sample and in the majority of single tumour cells; (2) subclonal mutations, detected in two or more single cells, but not in the bulk tumour; and (3) *de novo* mutations, found in only one tumour cell. The *de novo* mutations are difficult to distinguish from technical errors and were therefore excluded from our initial analysis. In total we detected 12 clonal non-synonymous mutations and 32 subclonal mutations (Fig. 2e). Many subclonal mutations occurred in intergenic regions; however, two mutations (*MARCH11* and *CABP2*) were found in coding regions (Supplementary Table 4).

To identify additional subclonal mutations, we performed single nuclei exome sequencing on a larger set of cells (47 tumour cells and 12 normal cells). Each nucleus was sequenced at 46.78× (46.78 ± 4.95, s.e.m., $n = 59$) coverage depth and 92.77% (92.77 ± 4.85, s.e.m., $n = 59$) coverage breadth, from which somatic mutations were detected (Supplementary Table 5). The mutations were clustered and sorted by frequency to construct a heatmap (Fig. 2f). As expected, the 17 clonal mutations identified by population sequencing were present in many of the single tumour cells, however, we also identified 22 new subclonal mutations. In contrast, only a single subclonal mutation was detected in the 12 normal cells (Fig. 2f, right panel).

Population and single nuclei sequencing of a TNBC

We then proceeded to analyse a triple-negative (ER$^-$/PR$^-$/Her2$^-$) breast cancer (TNBC) (Fig. 3a). We performed population sequencing of the bulk tumour (72×) and matched normal tissue (74×), and identified 374 non-synonymous mutations. A number of mutations occurred in cancer genes, including *PTEN*, *TBX3*, *NOTCH2*, *JAK1*, *ARAF*, *NOTCH3*,

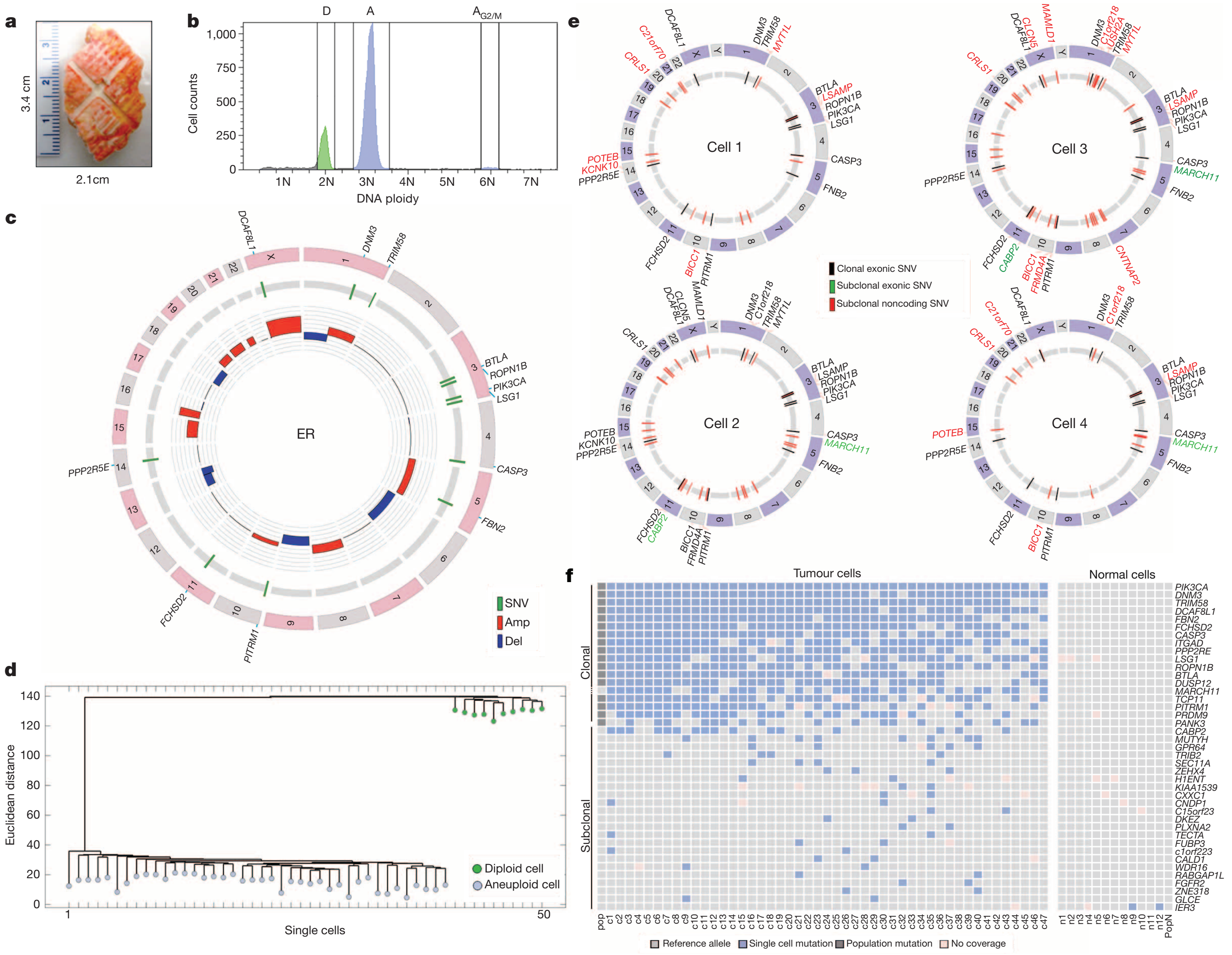


Figure 2 | Single cell and population sequencing of an ER tumour. a, Frozen ER tumour specimen. **b**, Flow-sorting histogram of ploidy distributions. **c**, Circos plot of mutations and CNAs detected in the population of aneuploid tumour cells. Cancer genes are on the outer ring. **d**, Neighbour-joining tree of integer copy number profiles from single diploid and aneuploid cells, rooted by the diploid node. **e**, Circos plots of whole-genome single cell sequencing data showing mutations detected in two or more cells. **f**, Heatmap of coding mutations detected by single-nuclei exome sequencing. Mutations detected by whole-genome sequencing (pop) and exome sequencing (ex) are also displayed.

MAP3K4, NTRK1, AFF4, CDH6, SETBP1, AKAP9, MAP2K7, ECM2 and *ECM1* (Supplementary Table 6) (Fig. 3b). Many of these mutations were previously reported in the TCGA breast cancer cohort[7]. Pathway analysis revealed two major pathways that were disrupted during tumour evolution: TGF-β ($P = 9.9 \times 10^{-2}$) and extracellular matrix-receptor signalling ($P = 2.7 \times 10^{-2}$). Copy number profiling identified many chromosomal deletions, in addition to a focal amplification on chromosome 19p13.2 (Fig. 3b).

To investigate genomic diversity at single cell resolution, we performed copy number profiling and exome sequencing. We flow-sorted 50 single nuclei from the hypodiploid (H), diploid (D) and aneuploid (A) ploidy distributions for copy number profiling using SNS (Fig. 3c). Neighbour-joining revealed two distinct subpopulations of tumour cells (A and H) in addition to the normal diploid cells (Fig. 3d). The single cell copy number profiles were analysed using clustered heatmaps, which showed highly similar rearrangements within each subpopulation (A mean $R^2 = 0.91$, H mean $R^2 = 0.88$), but were distinguished by two large deletions on chromosome 9 and 15 (Extended Data Fig. 3b).

Next, we flow-sorted 16 single tumour nuclei from the G2/M peaks (H and A) and 16 single normal nuclei for exome sequencing using nuc-seq (Fig. 3e). Non-synonymous point mutations were used to perform hierarchical clustering and multi-dimensional scaling (MDS). As expected, the 374 clonal non-synonymous mutations detected by bulk sequencing were found in the majority of the single tumour cells, however, we also identified 145 additional subclonal non-synonymous mutations that were not detected in the bulk tumour (Supplementary Table 7). MDS identified 4 distinct clusters, corresponding to three tumour subpopulations (H, A_1 and A_2) and the normal cells (Extended Data Fig. 5a). Hierarchical clustering showed that many of the subclonal mutations occurred exclusively in one subpopulation (H, A_1 or A_2) (Fig. 3e). The A_1 subpopulation contained 66 unique subclonal non-synonymous mutations, including *AURKA, SYNE2* and *PPP2R1A*. The A_2 subpopulation contained 52 unique subclonal non-synonymous mutations including *TGFB2* and *CHRM5*. In contrast only two subclonal mutations were shared between the normal cells (Fig. 3e, right panel). Many of the subclonal mutations (23.44%) were predicted to damage protein function by both POLYPHEN[21] and SIFT[22] (Extended Data Fig. 5b).

Single-molecule targeted deep sequencing

To validate the mutations detected by single cell sequencing and determine their frequencies in the bulk tumour, we performed targeted single-molecule deep-sequencing. Duplex libraries were constructed

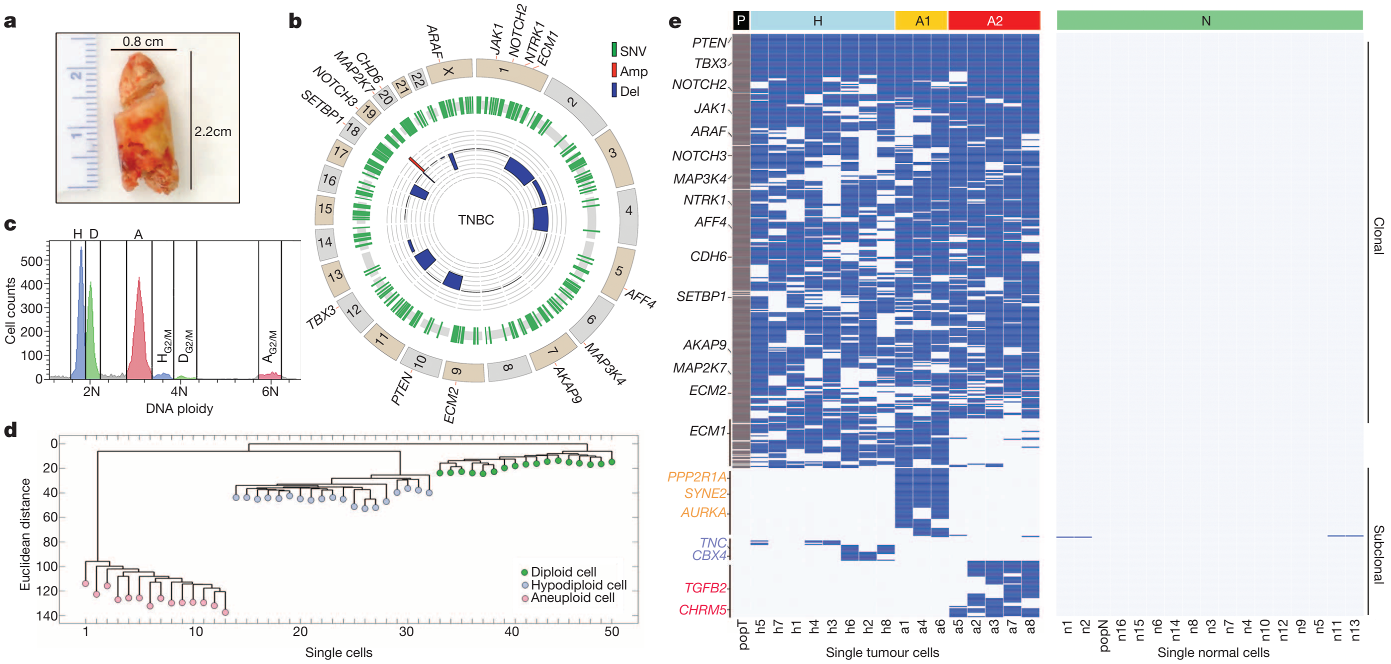


Figure 3 | Single cell and population sequencing of a triple-negative breast cancer. a, Frozen TNBC specimen. **b**, Circos plot of mutations and CNAs detected by population sequencing of the TNBC, with cancer genes on the outer ring. **c**, Flow-sorting histogram of ploidy distributions, showing three major subpopulations: diploid (D), hypodiploid (H) and aneuploid (A). **d**, Neighbour-joining tree of 50 single cell integer copy number profiles, rooted by the diploid node. **e**, Clustered heatmap of the nonsynonymous point mutations detected by single nuclei exome sequencing and population sequencing (P). Mutations detected in one cell are excluded.

from bulk tissue to reduce the error rate of next-generation sequencing[10]. Custom capture platforms were designed to target mutations detected in the single cells of the ERBC and TNBC tumours (Methods). Targeted deep-sequencing (116,952×) was performed in the ER tumour resulting in a single-molecule coverage depth of 5,695× using single-strand consensus sequences (SSCS). Deep-sequencing of the TNBC (118,743×) resulted in a single-molecule coverage depth of 6,634× using SSCS (Extended Data Fig. 4). We found that 61.5% of the reads were in the target regions in the ERBC and 80.2% in the TNBC.

The ERBC duplex data validated 94.44% (17/18) of the clonal mutations, 90.47% (19/21) of the subclonal mutations, and 19.40% (26/134) of the *de novo* mutations detected by single cell sequencing ($P < 0.01$) (Methods). The clonal mutations occurred at high frequencies in the tumour mass, whereas the subclonal mutations (0.0895 mean) and *de novo* mutations (0.0195 mean) were very rare (Fig. 4a). Similarly, in the TNBC we validated 99.73% (374/375) of the clonal mutations, 64.83% (94/145) of the subclonal mutations and 26.99% (152/563) of the *de novo* mutations ($P < 0.01$) (Methods). Similarly, we found that the clonal mutations in the TNBC showed high frequencies (0.4457 mean), however, the subclonal mutations were less prevalent (0.050 mean) and the *de novo* mutations were very rare (0.00047 mean) (Fig. 4b). This data suggests that many of the subclonal and *de novo* mutations are likely to be real biological variants that occur at low frequencies in the tumour mass.

Mathematical modelling of the mutation rates

To estimate the mutation rates in each tumour, we used the single cell mutation frequencies and designed a mathematical stochastic birth-and-death process model that uses experimentally derived parameters for cell birth rates (Ki-67 staining), cell death rates (caspase-3 staining), total tumour cell numbers (flow-sorting cell counts) and the tumour mass doubling time for invasive carcinomas (mean = 168 days)[23–25] (Methods). We modelled data for a series of mutation rates and compared the data to the empirical single cell mutation frequency distributions (Supplementary Table 8). Our data suggest that the ERBC had a mutation rate of $M_R = 0.6$ mutations per cell division for the exome data (Fig. 4c) and $M_R = 0.9$ for the single cell whole-genome data (Fig. 4d). These data are similar to the error rates reported for normal cells, which are approximately 0.6 mutations per cell division (error rate $= 1 \times 10^{-10}$)[26–28]. In contrast, our modelling suggests a mutation rate of $M_R = 8$ for the TNBC, suggesting a 13.3× fold increase relative to normal cells (Fig. 4e).

Discussion

In this study we report the development of a novel single cell genome sequencing method that utilizes G2/M nuclei to achieve high-coverage data with low error rates. Although G2/M nuclei were used in this study, the experimental protocol can also be used to sequence nuclei at any stage of the cell cycle. We applied nuc-seq to delineate clonal diversity and investigate mutational evolution in two breast cancer patients. Our data clearly show that no two single tumour cells are genetically identical, calling into question the strict definition of a clone. In both patients we observed a large number of subclonal and *de novo* mutations. These data suggest that point mutations evolved gradually over long periods of time, generating extensive clonal diversity (Fig. 4f, g). In contrast, the single cell copy number profiles were highly similar, suggesting that chromosome rearrangements occurred early, in punctuated bursts of evolution, followed by stable clonal expansions to form the tumour mass (Fig. 4h, i).

We previously reported punctuated copy number evolution by sequencing single cells from a TNBC patient[11]. This model has also been supported by bulk sequencing data in prostate cancer[29] and in rearrangement patterns called firestorms[30] or chromothripsis[31]. A punctuated model is consistent with the mechanisms that underlie CNAs, including chromosome missegregation[32], cytokinesis defects and breakage-fusion-bridge[33], which can generate complex rearrangements in just a few cell divisions. In contrast, point mutations occur through defects in DNA repair or replication machinery[34], which accumulate more gradually over many cell divisions. Our data are consistent with these mechanisms, and further show that two distinct molecular clocks were operating at different stages of tumour growth (Extended Data Fig. 6).

A pervasive problem in the field of single cell genomics is the inability to validate mutations that are detected in single cells. To address this problem, we combined single cell sequencing with targeted single-molecule

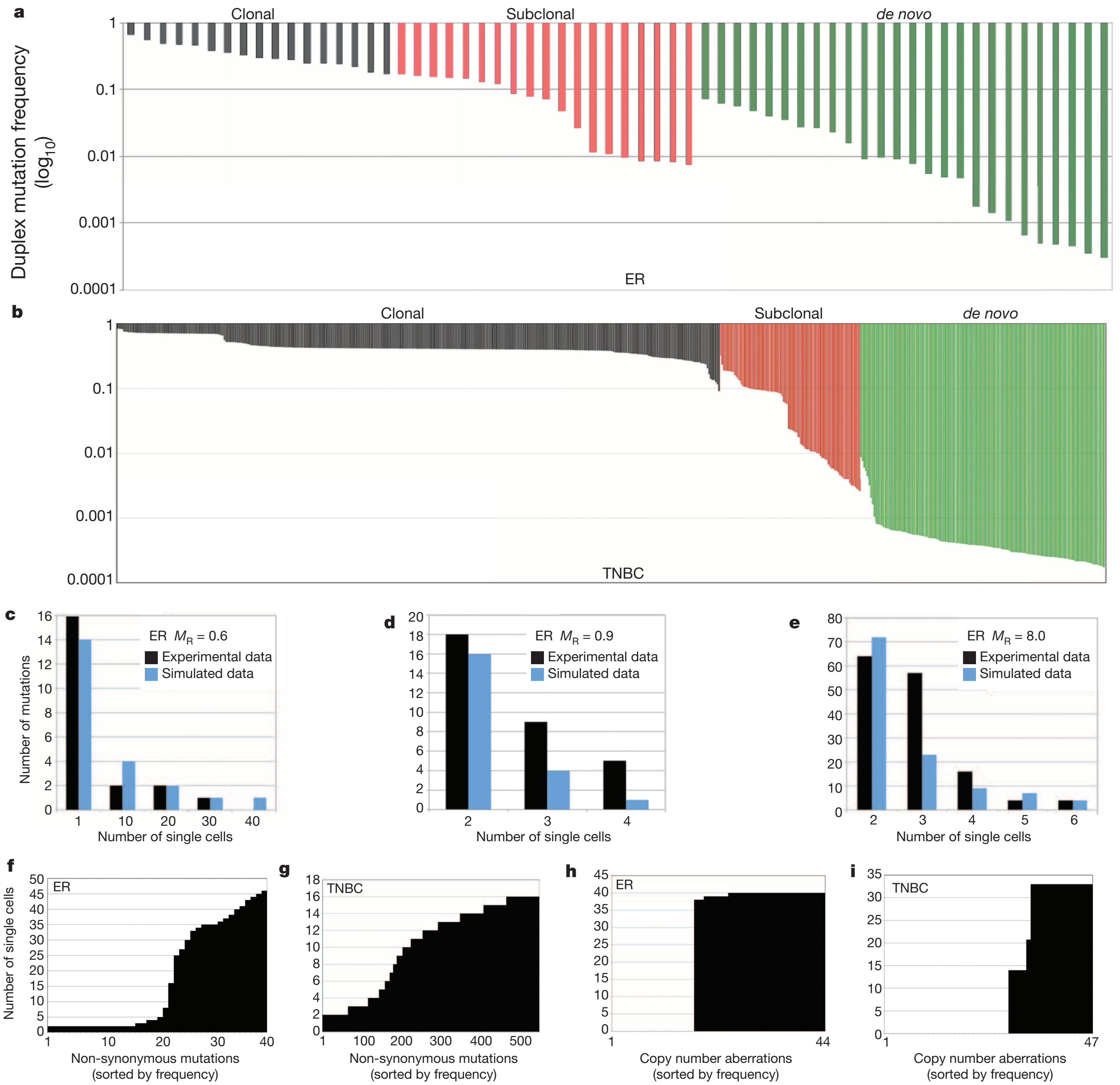


Figure 4 | Duplex mutation frequencies and mutation rates. a, ERBC duplex mutation frequencies from targeted deep-sequencing of the bulk tumour tissue. **b**, TNBC duplex mutation frequencies from deep-sequencing of the bulk tumour tissue. **c–e**, Mathematical modelling of mutation rates compared to experimental data. **c**, ERBC single-nuclei exome and modelling data at 0.6 mutation rate. **d**, ERBC whole-genome single nuclei and modelling data at 0.9 mutation rate. **e**, TNBC single nuclei exome and modelling data at a mutation rate of 8. **f**, Mutation frequencies shared by 2 or more cells in the ERBC. **g**, Mutation frequencies shared by 2 or more cells in the TNBC. **h**, CNAs shared by two or more cells in the ERBC. **i**, CNAs shared by two or more cells in the TNBC.

deep-sequencing. This approach not only validates mutations, but also measures the precise mutation frequencies in the bulk population. Using this approach, we identified hundreds of subclonal and *de novo* mutations that were present at low frequencies (<10%) in the tumour mass. These rare mutations may have an important role in diversifying the phenotypes of cancer cells, allowing them to survive selective pressures in the tumour microenvironment, including the immune system, hypoxia and chemotherapy[35,36].

A salient question in the field of chemotherapy is whether resistance mutations are pre-existing in rare cells in the tumour, or alternatively, emerge spontaneously in response to being challenged by the therapeutic agent. Although this question has been studied for decades in bacteria[37], it remains poorly understood in human cancers. Our data suggest that a large number of diverse mutations are likely to be pre-existing in the tumour mass before chemotherapy. Our data also has important implications for the mutator phenotype, which posits that tumour evolution is driven by increased mutation rates[34,38]. Although TCGA studies[39–41] report increased mutation frequencies, it remains unclear whether these mutations accumulate over many cell divisions (at a normal error rate) or through an increased mutation rate. Our TNBC data suggest an increased mutation rate (13.3×) relative to the normal cells, supporting this model.

We expect that single cell genome sequencing will open up new avenues of investigation in many diverse fields of biology. In cancer research there will be immediate applications for studying cancer stem cells and circulating tumour cells. In the clinic, these tools will have important applications in early detection and non-invasive monitoring. Beyond cancer, these tools will have utility in microbiology, development, immunology and neuroscience and will lead to substantial improvements in our fundamental understanding of human diseases.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 25 July 2012; accepted 23 June 2014.

Published online 30 July 2014.

1. Torres, L. *et al.* Intratumor genomic heterogeneity in breast cancer with clonal divergence between primary carcinomas and lymph node metastases. *Breast Cancer Res. Treat.* **102,** 143–155 (2007).
2. Navin, N. *et al.* Inferring tumor progression from genomic heterogeneity. *Genome Res.* **20,** 68–80 (2010).
3. Park, S. Y., Gonen, M., Kim, H. J., Michor, F. & Polyak, K. Cellular and genetic diversity in the progression of *in situ* human breast carcinomas to an invasive phenotype. *J. Clin. Invest.* **120,** 636–644 (2010).
4. Sørlie, T. *et al.* Gene expression patterns of carcinomas distinguish tumor subclasses with clinical implications. *Proc. Natl Acad. Sci. USA* **98,** 10869–10874 (2001).
5. Curtis, C. *et al.* The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* **486,** 346–352 (2012).
6. Shah, S. P. *et al.* The clonal and mutational evolution spectrum of primary triple-negative breast cancers. *Nature* **486,** 395–399 (2012).
7. The Cancer Genome Atlas Network Comprehensive molecular portraits of human breast tumours. *Nature* **490,** 61–70 (2012).
8. Nik-Zainal, S. *et al.* The life history of 21 breast cancers. *Cell* **149,** 994–1007 (2012).
9. Ellis, M. J. *et al.* Whole-genome analysis informs breast cancer response to aromatase inhibition. *Nature* **486,** 353–360 (2012).
10. Schmitt, M. W. *et al.* Detection of ultra-rare mutations by next-generation sequencing. *Proc. Natl Acad. Sci. USA* **109,** 14508–14513 (2012).
11. Navin, N. *et al.* Tumour evolution inferred by single-cell sequencing. *Nature* **472,** 90–94 (2011).
12. Woyke, T. *et al.* One bacterial cell, one complete genome. *PLoS ONE* **5,** e10314 (2010).
13. Dichosa, A. E. *et al.* Artificial polyploidy improves bacterial single cell genome recovery. *PLoS ONE* **7,** e37387 (2012).
14. Hou, Y. *et al.* Single-cell exome sequencing and monoclonal evolution of a *JAK2*-negative myeloproliferative neoplasm. *Cell* **148,** 873–885 (2012).
15. Klein, C. A. *et al.* Comparative genomic hybridization, loss of heterozygosity, and DNA sequence analysis of single cells. *Proc. Natl Acad. Sci. USA* **96,** 4494–4499 (1999).
16. Adey, A. *et al.* Rapid, low-input, low-bias construction of shotgun fragment libraries by high-density *in vitro* transposition. *Genome Biol.* **11,** R119 (2010).
17. Kytola, S. *et al.* Chromosomal alterations in 15 breast cancer cell lines by comparative genomic hybridization and spectral karyotyping. *Genes Chromosomes Cancer* **28,** 308–317 (2000).
18. Baslan, T. *et al.* Genome-wide copy number analysis of single cells. *Nature Protocols* **7,** 1024–1041 (2012).
19. Zong, C., Lu, S., Chapman, A. R. & Xie, X. S. Genome-wide detection of single-nucleotide and copy-number variations of a single human cell. *Science* **338,** 1622–1626 (2012).
20. Lorenz, M. O. Methods of measuring the concentration of wealth. *J. Am. Stat. Assoc.* **9,** 209–219 (1905).
21. Adzhubei, I. A. *et al.* A method and server for predicting damaging missense mutations. *Nature Methods* **7,** 248–249 (2010).
22. Ng, P. C. & Henikoff, S. SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res.* **31,** 3812–3814 (2003).
23. Kuroishi, T. *et al.* Tumor growth rate and prognosis of breast cancer mainly detected by mass screening. *Jpn. J. Cancer Res.* **81,** 454–462 (1990).
24. Peer, P. G., van Dijck, J. A., Hendriks, J. H., Holland, R. & Verbeek, A. L. Age-dependent growth rate of primary breast cancer. *Cancer* **71,** 3547–3551 (1993).
25. Michaelson, J. *et al.* Estimates of breast cancer growth rate and sojourn time from screening database information. *J. Women's Imaging* **5,** 11–19 (2003).
26. Nachman, M. W. & Crowell, S. L. Estimate of the mutation rate per nucleotide in humans. *Genetics* **156,** 297–304 (2000).
27. Drake, J. W., Charlesworth, B., Charlesworth, D. & Crow, J. F. Rates of spontaneous mutation. *Genetics* **148,** 1667–1686 (1998).
28. Preston, B. D., Albertson, T. M. & Herr, A. J. DNA replication fidelity and cancer. *Semin. Cancer Biol.* **20,** 281–293 (2010).
29. Baca, S. C. *et al.* Punctuated evolution of prostate cancer genomes. *Cell* **153,** 666–677 (2013).
30. Hicks, J. *et al.* Novel patterns of genome rearrangement and their association with survival in breast cancer. *Genome Res.* **16,** 1465–1479 (2006).
31. Stephens, P. J. *et al.* Massive genomic rearrangement acquired in a single catastrophic event during cancer development. *Cell* **144,** 27–40 (2011).
32. Pellman, D. Cell biology: aneuploidy and cancer. *Nature* **446,** 38–39 (2007).
33. McClintock, B. The stability of broken ends of chromosomes in *Zea mays*. *Genetics* **26,** 234–282 (1941).
34. Loeb, L. A. Human cancers express mutator phenotypes: origin, consequences and targeting. *Nature Rev. Cancer* **11,** 450–457 (2011).
35. Merlo, L. M. F., Pepper, J. W., Reid, B. J. & Maley, C. C. Cancer as an evolutionary and ecological process. *Nature Rev. Cancer* **6,** 924–935 (2006).
36. Greaves, M. & Maley, C. C. Clonal evolution in cancer. *Nature* **481,** 306–313 (2012).
37. Luria, S. E. & Delbruck, M. Mutations of bacteria from virus sensitivity to virus resistance. *Genetics* **28,** 491–511 (1943).
38. Bielas, J. H., Loeb, K. R., Rubin, B. P., True, L. D. & Loeb, L. A. Human cancers express a mutator phenotype. *Proc. Natl Acad. Sci. USA* **103,** 18238–18242 (2006).
39. Lawrence, M. S. *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499,** 214–218 (2013).
40. Alexandrov, L. B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500,** 415–421 (2013).
41. Kandoth, C. *et al.* Mutational landscape and significance across 12 major cancer types. *Nature* **502,** 333–339 (2013).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank L. Ramagli, H. Tang, E. Thompson, K. Khanna, W. Schober and J. Tyler. We are grateful to S. Kennedy and L. Loeb for help with the duplex protocols. We thank M. Edgerton, J. Hicks, M. Wigler and J. Kendall for discussions. We thank R. Krahe and M. Rui for reviewing the manuscript. N.E.N. is a Nadia's Gift Foundation Damon Runyon-Rachleff Innovator (DRR-25-13). This research was supported by grants to N.E.N. from NIH (R21CA174397-01) and NCI (1RO1CA169244-01). N.E.N. was supported by T.C. Hsu and the Alice-Reynolds Kleberg Foundation. N.E.N. and P.S. were supported by the Center for Genetics & Genomics. F.M.-B was supported by an NIH UL1 (TR000371) and Susan Komen (SAC10006). K.C. was supported by the NCI (RO1CA172652). H.L. was supported by the NIH (U24CA143883). F.M. was supported by PS-OC (U54CA143798). K.C. and H.L. were supported by the Dell Foundation. M.L.L. is a CPRIT scholar and is supported by ALA. This work was also supported by an NCI center grant (CA016672). A.U. is a Rosalie B. Hite Fellow.

Author Contributions Y.W. performed experiments and data analysis. M.L.L., J.W., A.M. and X.S. performed experiments. A.U., W.R., K.C., H.L., P.S. and S.V. performed data and statistical analyses. H.Z. and F.M.-B. obtained clinical samples. R.Z. and F.M. performed modelling. N.E.N. performed experiments, analysed data and wrote the manuscript.

Author Information The data from this study has been deposited into the Sequence Read Archive (SRA053195). Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to N.E.N. (nnavin@mdanderson.org).

doi:10.1038/nature13566

Crystal structure of the human COP9 signalosome

Gondichatnahalli M. Lingaraju[1,2]*, Richard D. Bunker[1,2]*, Simone Cavadini[1,2], Daniel Hess[1], Ulrich Hassiepen[3], Martin Renatus[3], Eric S. Fischer[1,2] & Nicolas H. Thomä[1,2]

Ubiquitination is a crucial cellular signalling process, and is controlled on multiple levels. Cullin–RING E3 ubiquitin ligases (CRLs) are regulated by the eight-subunit COP9 signalosome (CSN). CSN inactivates CRLs by removing their covalently attached activator, NEDD8. NEDD8 cleavage by CSN is catalysed by CSN5, a Zn^{2+}-dependent isopeptidase that is inactive in isolation. Here we present the crystal structure of the entire ~350-kDa human CSN holoenzyme at 3.8 Å resolution, detailing the molecular architecture of the complex. CSN has two organizational centres: a horseshoe-shaped ring created by its six proteasome lid–CSN–initiation factor 3 (PCI) domain proteins, and a large bundle formed by the carboxy-terminal α-helices of every subunit. CSN5 and its dimerization partner, CSN6, are intricately embedded at the core of the helical bundle. In the substrate-free holoenzyme, CSN5 is autoinhibited, which precludes access to the active site. We find that neddylated CRL binding to CSN is sensed by CSN4, and communicated to CSN5 with the assistance of CSN6, resulting in activation of the deneddylase.

CSN, which was first discovered in *Arabidopsis thaliana* as a repressor of constitutive photomorphogenesis[1], is a protein complex common to all eukaryotes[2]. CSN regulates CRLs[3–5], a family of ~200 complexes in humans implicated in many regulatory processes[6] that together direct ~20% of proteasome-mediated protein degradation[7]. Enzymatically, CSN functions as an isopeptidase that removes the ubiquitin-like activator NEDD8 from CRLs[8], but it can also bind deneddylated CRLs and maintain them in an inactive state[9–11]. CRLs are composed of a cullin protein backbone on which a RING domain-containing protein (RBX1 or RBX2) and a substrate receptor module are bound[12]. Ubiquitin-loaded E2 ubiquitin conjugating enzymes are recruited by CRLs using RBX1 or RBX2 to ubiquitinate substrates recognized by their receptor. CRL activity is stimulated by the conjugation of NEDD8 to a conserved lysine residue in the cullin C-terminal domain[13–17]. CSN has emerged as the sole enzyme capable of removing NEDD8 modifications from cullins with proficiency[3,4,8,11]. It is exquisitely specific for neddylated CRLs and, unlike other isopeptidases, has neither general deubiquitination nor deneddylation activity[8,9].

Human CSN contains eight distinct proteins (designated CSN1–8 by decreasing molecular weight, from 57 to 22 kDa), all of which are required for full enzymatic activity *in vitro*[18]. The essentiality of CSN has been demonstrated in model organisms: in *A. thaliana*, the loss of any subunit is lethal in the development of seedlings[19]; similarly, in mice, knockout of CSN2, 3, 5, 6 or 8 is lethal at the embryonic stage[19,20]. Several CSN subunits in humans show elevated expression in cancer and have been implicated in sustaining oncogenic transformation[21].

CSN5 provides the catalytic centre for CSN, yet as an isopeptidase it is essentially inactive outside the holoenzyme[8,18,22]. This raises intriguing questions as to how CSN5 is harnessed by CSN and the nature of the regulatory mechanism that imposes strict substrate specificity.

Knowledge of the structure of CSN has been limited to low-resolution electron microscopy models of CSN alone and in complex with CRL1 family members[10,23], and those obtained for the related complexes, the 19S lid component of 26S proteasome (19S lid)[24–28] and the eukaryotic translation initiation factor (eIF3)[29–32]. Interpretation of these maps has been aided by molecular models, which have been determined for parts of a few individual proteins[22,33–41]. Detailed structural studies of CSN, required for understanding its unique activity and selectivity towards neddylated CRLs have, however, posed a challenge.

The 3.8 Å resolution CSN crystal structure presented here provides detailed insight into the molecular architecture of the complex. We find CSN captured in the crystal in an inactive state, wherein CSN5 occludes its own active site. Binding of a neddylated CRL to the holoenzyme triggers substantial remodelling of CSN4, 5 and 6, resulting in activation of the CSN5 isopeptidase.

Structure of CSN

Human CSN, consisting of CSN1, 2, 3, 4, 5, 6, 7a and 8, was co-expressed and purified from insect cells (Methods). Crystals of CSN lacking flexible regions of CSN1 (isoform 2, 1–51), CSN5 (residues 1–11) and CSN7a (residues 219–275) were obtained and their structure determined by X-ray crystallography. These truncations did not impair holoenzyme formation or catalytic activity (Extended Data Fig. 1a–d, l). The final model includes the two CSN complexes found in the asymmetric unit of the crystals, 5,178 amino acids in 16 protomers and two Zn^{2+} ions (Extended Data Table 1a and Extended Data Figs 2a–g, 3a–l). Six CSN proteins (CSN1–4 and CSN7–8) contain a PCI domain, characterized by helical repeats followed by a winged-helix (WH) subdomain[34,42,43] (Fig. 1a–d). The other two subunits, CSN5 and CSN6, have MPR1/PAD1 amino-terminal (MPN) domains, a metalloprotease fold[22,36,44,45]. Only CSN5, however, has a complete active site and binds zinc. All of the subunits have C-terminal helical decorations separated by largely structured linkers from their core domains (Supplementary Data).

CSN has overall dimensions of 173 × 142 × 108 Å (Fig. 1a–c). The complex is governed from two organizational centres (Fig. 1e): an open ring formed by association of the WH subdomains from the six PCI proteins (PCI ring), and an elaborate bundle comprising the C-terminal α-helices from each subunit (helical bundle). The PCI and MPN proteins

[1] Friedrich Miescher Institute for Biomedical Research, Maulbeerstrasse 66, 4058 Basel, Switzerland. [2] University of Basel, Petersplatz 10, 4003 Basel, Switzerland. [3] Novartis Pharma AG, Institutes for Biomedical Research, Novartis Campus, 4056 Basel, Switzerland.
*These authors contributed equally to this work.

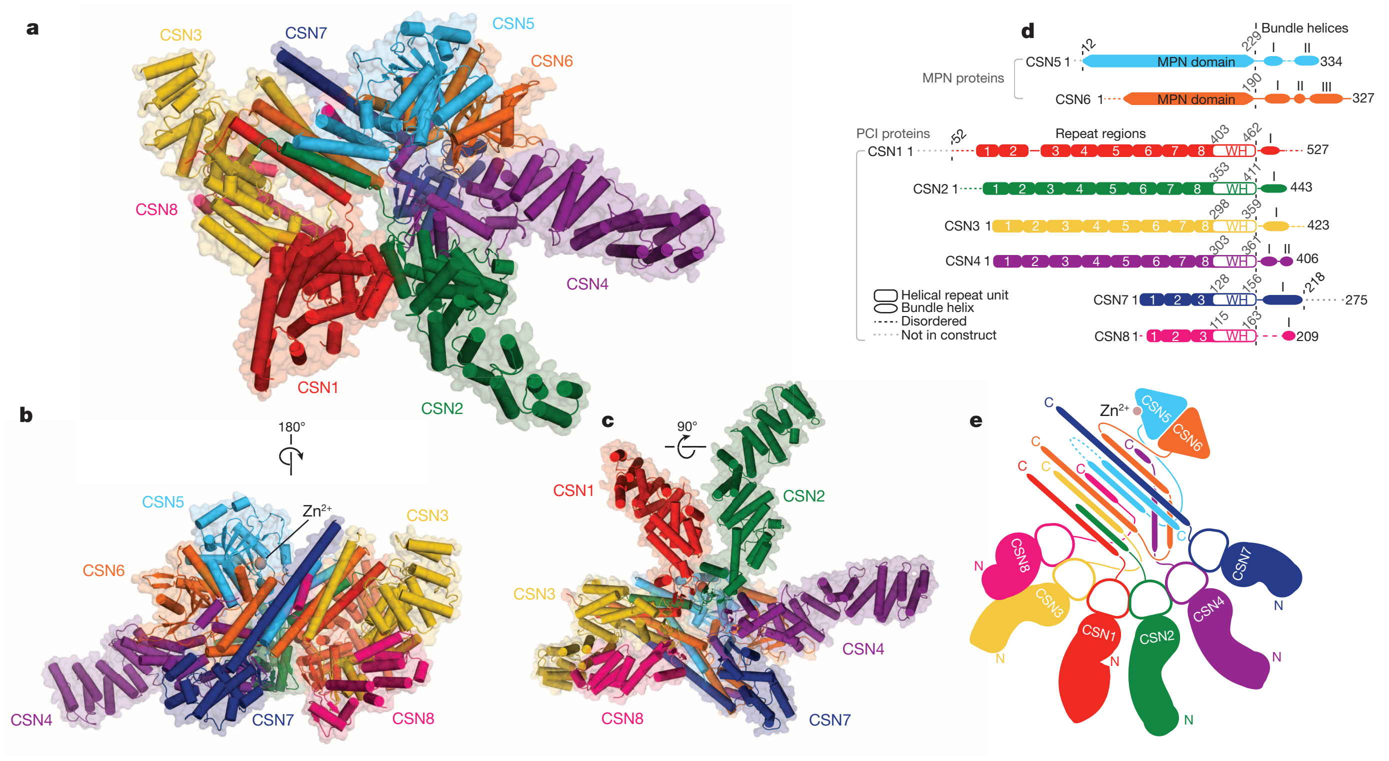


Figure 1 | **Overall architecture of CSN.** **a–c**, Cartoon representation of CSN in three orientations. **d**, A schematic representation of the domain organization of the CSN proteins. Domain boundaries are indicated. **e**, A flattened schematic representation of the three-dimensional structure of CSN. The WH subdomains forming the PCI ring are shown as white rings. Beyond the two organizational centres and the CSN5–CSN6 dimer, interactions are formed between CSN3 and the CSN8 N-terminal repeats, the CSN7 N-terminal repeat and the helical bundle (see Extended Data Fig. 5f, g).

form largely distinct subassemblies that are united in the helical bundle. The N-terminal helical repeat domains of the PCI proteins (CSN1–4 and CSN7–8) radiate from the PCI ring at the base of the complex, the largest of which (CSN1–4) form prominent arm-like protrusions (Fig. 1a, e). The helical bundle sits across the PCI ring. A heterodimer formed by the MPN domains of CSN5 and CSN6 rests on the helical bundle (Fig. 1a, e). The MPN dimer (CSN5–CSN6 dimer), helical bundle and PCI ring create an intricate three-layered assembly. This overall architecture is shared among CSN and its paralogous complexes, the 19S lid and eIF3 (Extended Data Fig. 4a–h and Supplementary Discussion).

Detail of the two organizational centres

The PCI proteins are organized about an open ring formed by association of their WH subdomains (Fig. 2a). The short three-stranded β-sheets in each WH subdomain are oligomerized edge-to-edge in the order CSN7–CSN4–CSN2–CSN1–CSN3–CSN8 to form an 18-stranded composite β-sheet at the centre of the complex (Fig. 2b–d). The central β-sheet has right-handed curvature and progresses through one incomplete helical turn (~300°), resulting in a horseshoe-shaped appearance when viewed down the β-strand axis (Fig. 2b, c, Extended Data Fig. 5a–g and Supplementary Data).

The helical bundle lies over the PCI ring at an angle of ~110° from the plane of the β-sheet (Fig. 3a). CSN6 forms a U-shaped structure at the centre of the bundle with its three C-terminal helices (helices I–III) that interacts with every other subunit (Fig. 3b, c). The helices from CSN1, 2, 3 and 8 surround CSN6 helix III. The 80-Å-long helix from CSN7 (helix I) contacts CSN6 helices I and II at the base of the bundle, nearest the PCI ring. The two helices from CSN4 (helices I and II) form a brace roughly perpendicular to the bundle axis in contact with the three C-terminal helices of CSN6. CSN5, whose two C-terminal helices form an antiparallel hairpin, inserts its final C-terminal helix (helix II) into the central CSN6 framework at the core of the bundle. Deletion of the C-terminal helices has a pronounced effect on CSN integrity[46] (Extended Data Fig. 6a–g and Supplementary Data).

CSN5–CSN6 heterodimer

The MPN domains of CSN5 and 6 form an intimate dimer with pseudo-two-fold symmetry using an interface that buries ~900 $\text{\AA}^2$ of surface area (Fig. 4a, b). Although CSN5 and 6 share sequence and structural similarity, the catalytic and Zn^{2+}-coordinating residues are only present in CSN5 (Fig. 4b, c; see later). The CSN5 and 6 chains, which are

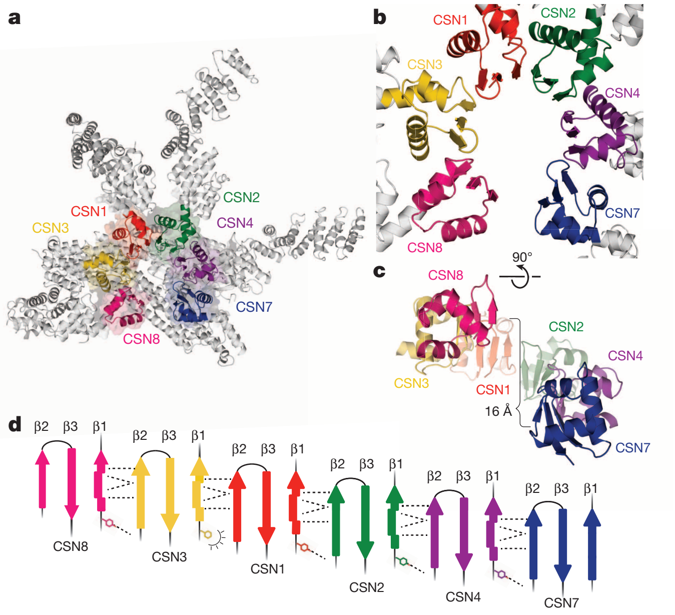


Figure 2 | **PCI ring assembly.** **a**, Cartoon representation of CSN (grey) with the PCI ring highlighted in colour. **b**, Close-up of **a**. **c**, Alternative view of **b**, illustrating the opening of the ring. **d**, Schematic representation of the composite β-sheet. Recurrent hydrogen bonding interactions between WH units are shown with dashed lines.

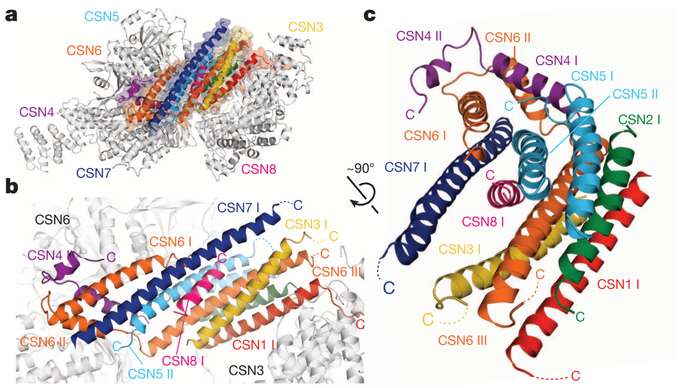


Figure 3 | Helical bundle assembly. a, Cartoon representation of CSN (grey) highlighting the helical bundle formed by the C-terminal helices of every subunit in colour. **b**, Close-up of **a**. The C-terminal helices are numbered with roman numerals as in Fig. 1d. Disordered residues are represented by dashed lines. **c**, An alternative view of **b**.

topologically knotted, cross over each other before entering the helical bundle (Extended Data Fig. 6h), where they are also closely associated. Deletion of the CSN6 MPN domain, leaving its C-terminal helices to maintain complex integrity (Extended Data Fig. 6g), results in a CSN mutant that is severely catalytically impaired, exhibiting a 100-fold decrease in the turnover rate constant (k_{cat}) relative to wild-type CSN (Extended Data Fig. 1e, l). These observations point to a role for the CSN6 MPN domain in stabilizing the structure of the CSN5 MPN domain.

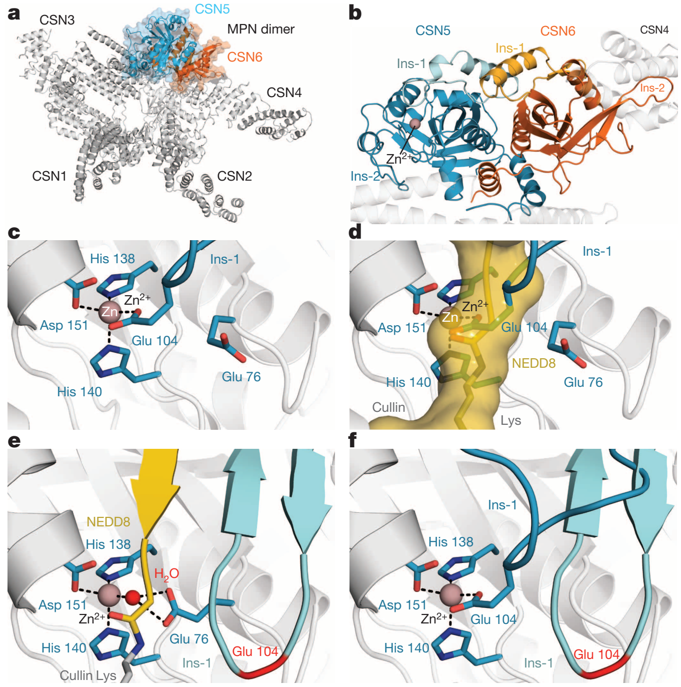


Figure 4 | CSN5 autoinhibition within CSN. a, Cartoon representation of CSN (grey) highlighting the CSN5–CSN6 dimer. **b**, Close-up of **a**. **c**, CSN5 active site. **d**, Docking of an isopeptide-linked neddylated CRL (yellow) into the CSN5 active site. The CRL–NEDD8 isopeptide bond (Protein Data Bank (PDB) accession 3DQV[48]) was fitted in the CSN5 active site based on the di-ubiquitin-bound structure of AMSH-LP (PDB accession 2ZNV[47]). The side chain of Glu 104 coordinates the Zn^{2+} ion and blocks access to the active site, autoinhibiting CSN5. **e**, Model of a catalytically competent state of the CSN5 active centre based on AMSH-LP and thermolysin. **f**, Superposition of the Ins-1 loop from **c** and **e**.

Integration of CSN5 in CSN

CSN provides the essential framework for CSN5 to function as an isopeptidase. Despite its entangled structure, the order of CSN assembly seems to be surprisingly lenient. CSN5 can enter an otherwise complete seven-subunit CSN and yield a holoenzyme capable of catalysis (Extended Data Fig. 7a, b). Because catalytic activity is directly proportional to the fraction of CSN complexes that have CSN5 included, we asked whether the absence of a given subunit prevents CSN5 integration (Extended Data Fig. 7c, d). Although the absence of CSN8 or 3 was tolerated, excluding CSN1, 2, 4, 6 or 7 strongly disfavoured CSN5 incorporation. Assembly was completely disrupted by the omission of full-length CSN6, emphasizing its crucial structural role in the formation of the helical bundle.

CSN5 has an autoinhibited state in CSN

In CSN, the CSN5 active site Zn^{2+} ion is tetrahedrally coordinated by the side chains of His 138, His 140 and Asp 151 of the canonical JAB1/MPN/MOV34 (JAMM) motif, and the side chain of Glu 104 (Fig. 4c). Glu 104 is situated in the MPN domain insertion-1 loop segment (Ins-1), which is essential for substrate recognition in related isopeptidases. While Glu 104 is liganded to the Zn^{2+} ion, Ins-1 occludes the entire CSN5 active site (Fig. 4d). In CSN, CSN5 Glu 104 replaces the water molecule that acts as the nucleophile in the hydrolysis of the isopeptide bond in related isopeptidases[47]. This water is positioned and polarized by an essential acidic residue, which is Glu 76 in CSN5 (Fig. 4c). Mutating Glu 76 in CSN5 to Ala (CSN (CSN5(E76A))) inactivates CSN (Extended Data Fig. 1a). This is analogous to the inactivating mutation first described for the MPN domain protease from *Archaeoglobus fulgidus*, AfJAMM[45], suggesting a shared catalytic mechanism. Although the overall enzymatic mechanism appears to be conserved among MPN proteases, the active site of CSN5 in the crystal is not configured for catalysis (Fig. 4c, d). The Ins-1 conformation observed in the holoenzyme differs from that of the crystal structure of CSN5 alone (Extended Data Fig. 8a–c)[22] and other MPN proteases (AMSH-LP, RPN11)[40,41,47], which typically have the catalytic water coordinated to the active site Zn^{2+} ion. For CSN activation: (1) the Glu 104 ligand must be removed from the Zn^{2+} ion; (2) Ins-1 has to change conformation to position the substrate polypeptide; and (3) Glu 76 needs to orient towards the Zn^{2+} ion to activate the catalytic water (Fig. 4e, f; see Supplementary Discussion and Extended Data Fig. 8d–f). Hence, a mechanism must exist to trigger remodelling of CSN5 and activation of CSN. As discussed later, this conformational trigger appears to be binding of a neddylated CRL substrate.

Substrate-induced structural dynamics

To study how CSN interacts with a substrate, we examined the negative-stain electron microscopy structure of an activated CRL, ${}_{\text{N8}}$SCF$^{\text{SKP2/CKS1}}$ (NEDD8–CUL1–RBX1–SKP1–SKP2–CKS1) in complex with CSN (CSN–${}_{\text{N8}}$SCF$^{\text{SKP2/CKS1}}$)[10]. The CSN structure has good agreement with the CSN–${}_{\text{N8}}$SCF$^{\text{SKP2/CKS1}}$ electron microscopy map when fitted as a single rigid body (correlation coefficient of 0.77). Rigid body movement of the CSN4 helical repeats and the CSN5–CSN6 MPN dimer improved the fit (Fig. 5a, b). The model clearly shows the CSN subunits that contact ${}_{\text{N8}}$SCF$^{\text{SKP2/CKS1}}$, and reveals interactions extending beyond the localized interaction between CSN5 and the neddylated cullin. The concave face of CSN2 (helical repeats 2–5) embraces the CUL1 C-terminal arm (WH$_{\text{B}}$ domain) (Fig. 5a, b), similar to what has been proposed previously[10]. The SKP2–CKS1 substrate receptor is positioned within 10–20 Å of CSN3/CSN8. Comparing the CSN crystal structure with the CSN–${}_{\text{N8}}$SCF$^{\text{SKP2/CKS1}}$ electron microscopy model reveals a ~35° rotation of the CSN4 N-terminal helical repeats relative to its WH subdomain (Fig. 5a–c). The conformer in the high-resolution crystal structure of isolated CSN4, determined in the process of solving the CSN structure (Fig. 5c and Extended Data Fig. 9a, b), closely matches the CSN4 conformation observed in the CSN–${}_{\text{N8}}$SCF$^{\text{SKP2/CKS1}}$ electron microscopy map. The dramatic domain motion in CSN4 is enabled by a hinge loop at the end of the helical repeats (residues 291–298) (Extended Data Fig. 9a).

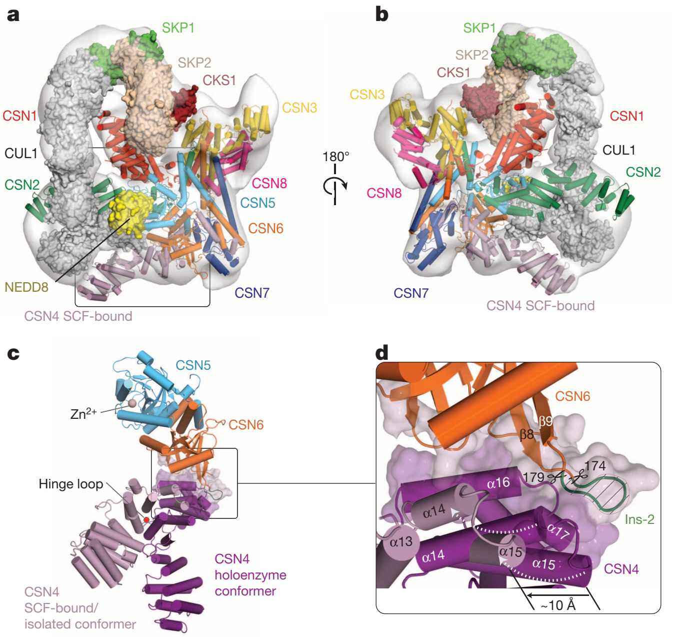


Figure 5 | The CSN–SCF interaction and CRL1-dependent CSN5 activation. **a**, **b**, Fit of crystallographic models of CSN and $_{\mathrm{N8}}\mathrm{SCF}^{\mathrm{SKP2/CKS1}}$ (PDB accessions 1LDK[49], 3DQV[48] and 2ASS[50]) into the CSN–$_{\mathrm{N8}}\mathrm{SCF}^{\mathrm{SKP2/CKS1}}$ electron microscopy map (Electron Microscopy Data Bank accession 2173 (ref. 10)). **c**, CSN4 conformations determined in isolation (mauve) (Extended Data Fig. 9a) and in the holoenzyme (purple) orientated as in **a** shown with CSN5–CSN6 dimer. **d**, Conformational changes in CSN4 are expected to impact the CSN5–CSN6 dimer. Close-up of the boxed region in **c** showing the portion of the CSN6 Ins-2 loop (green) in contact with CSN4. Scissors indicate the region removed in the CSN6$^{\Delta\mathrm{loop}}$ mutant (residues 174–179).

Finding evidence for substrate-induced conformational changes in CSN4 led us to examine its functional relevance. In the substrate-free CSN holoenzyme, CSN4 contacts the MPN domain of CSN6 through an extended interface involving a conserved β-hairpin loop (CSN6 residues 172–182; the insertion-2 (Ins-2) region in MPN domain proteins) (Fig. 5c, d). A conformational change in CSN4 following CRL binding would impact the CSN5–CSN6 dimer. Indeed, we observe considerable movement of the CSN5–CSN6 dimer in the CSN–$_{\mathrm{N8}}\mathrm{SCF}^{\mathrm{SKP2/CKS1}}$ electron microscopy envelope away from the helical bundle towards the neddylated cullin when compared with CSN in the crystal. A CSN6 deletion mutant lacking residues 174–179 of the Ins-2 loop that integrated stably into CSN (CSN (CSN6$^{\Delta\mathrm{loop}}$)) (Fig. 5d and Extended Data Fig. 1f) was used to probe the function of the CSN4–CSN6 interface. CSN (CSN6$^{\Delta\mathrm{loop}}$) had a k_{cat} 4.5-fold higher than wild-type CSN, but an indistinguishable Michaelis constant (K_{m}) value (Extended Data Fig. 1f, l). Thus, mutating the CSN4–CSN6 interface appears to remove an inhibitory component, yielding a complex more active than wild type.

In the absence of a bound neddylated CRL substrate, CSN maintains CSN5 in an autoinhibited state (Fig. 4c, d). Discovering a mutation in the CSN4–CSN6 interface that conferred greater activity to CSN prompted us to question whether the CSN4–CSN6 interface is part of a regulatory circuit that inhibits CSN5 until a neddylated CRL substrate is bound. Wild-type CSN has very limited isopeptidase activity when assayed with ubiquitin-rhodamine, a small artificial substrate for deubiquitinases (Extended Data Fig. 1h). CSN (CSN6$^{\Delta\mathrm{loop}}$), however, cleaves ubiquitin-rhodamine robustly with $k_{\mathrm{cat}} = 0.04\ \mathrm{s}^{-1}$; $K_{\mathrm{m}} = 1.8\ \mu\mathrm{M}$ (Extended Data Fig. 1i, m). Activity towards non-CRL substrates was also found for a CSN point mutant carrying CSN5 Glu104Ala in the CSN5 autoinhibitory loop (Ins-1) ($k_{\mathrm{cat}} = 0.04\ \mathrm{s}^{-1}$; $K_{\mathrm{m}} = 2.7\ \mu\mathrm{M}$) (Extended Data Fig. 1j, m). The double mutant combining CSN6$^{\Delta\mathrm{loop}}$ and CSN5 Glu104Ala, CSN (CSN6$^{\Delta\mathrm{loop}}$, CSN5(E104A)), had an additive effect on activity, producing a complex with greater catalytic activity on ubiquitin-rhodamine than wild type and either single mutant ($k_{\mathrm{cat}} = 0.2\ \mathrm{s}^{-1}$; $K_{\mathrm{m}} = 6.3\ \mu\mathrm{M}$) (Extended Data Fig. 1k, m). These results suggest that the CSN isopeptidase is inhibited by the CSN5 Ins-1 loop bearing Glu 104 and separately through the CSN4–CSN6 interface. The CRL substrate-induced conformational changes thus provide a mechanism by which non-CRLs are excluded from deneddylation.

We propose that binding of a neddylated CRL sensed by CSN4 facilitates movement of the CSN5–CSN6 dimer towards the neddylated CRL. The proximity of NEDD8 and the cullin to the CSN5 autoinhibitory Ins-1 loop may then be sufficient to remodel CSN5, leading to activation of CSN and deneddylation (Extended Data Fig. 9c).

Although CSN, eIF3 and the 19S lid share striking structural similarity, the intricate substrate-induced activation mechanism identified here seems to be unique to CSN[40,41] (see also Supplementary Discussion and Extended Data Fig. 4b–h).

Concluding remarks

The structural and functional characterization of CSN and analysis of its interaction with a neddylated CRL1 exposes functional roles for parts of the holoenzyme: the PCI ring (CSN1–4 and CSN7–8) organizes the helical repeat domains, which bind $_{\mathrm{N8}}\mathrm{SCF}^{\mathrm{SKP2/CKS1}}$ (principally through CSN2 (ref. 10)). The helical bundle enables CSN5 to sense the assembly state of CSN, favouring its own integration when the complex is otherwise fully assembled. Given that CSN5 is inactive in isolation, this ensures that the isopeptidase only becomes functional when CSN is equipped to bind CRLs and the induced fit mechanisms (provided by CSN4 and CSN6) are in place to activate CSN5 in response to a neddylated CRL. This interdependence of architecture and function explains why CSN acts exclusively on neddylated CRLs and avoids unregulated deubiquitinase activity.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 5 March; accepted 6 June 2014.
Published online 16 July 2014.

1. Wei, N., Chamovitz, D. A. & Deng, X. W. *Arabidopsis* COP9 is a component of a novel signaling complex mediating light control of development. *Cell* **78,** 117–124 (1994).
2. Wei, N. & Deng, X. W. The COP9 signalosome. *Annu. Rev. Cell Dev. Biol.* **19,** 261–286 (2003).
3. Lyapina, S. *et al.* Promotion of NEDD-CUL1 conjugate cleavage by COP9 signalosome. *Science* **292,** 1382–1385 (2001).
4. Schwechheimer, C. *et al.* Interactions of the COP9 signalosome with the E3 ubiquitin ligase SCF$^{\mathrm{TIR1}}$ in mediating auxin response. *Science* **292,** 1379–1382 (2001).
5. Zhou, C. *et al.* The fission yeast COP9/signalosome is involved in cullin modification by ubiquitin-related Ned8p. *BMC Biochem.* **2,** 7 (2001).
6. Lydeard, J. R., Schulman, B. A. & Harper, J. W. Building and remodelling Cullin–RING E3 ubiquitin ligases. *EMBO Rep.* **14,** 1050–1061 (2013).
7. Soucy, T. A. *et al.* An inhibitor of NEDD8-activating enzyme as a new approach to treat cancer. *Nature* **458,** 732–736 (2009).
8. Cope, G. A. *et al.* Role of predicted metalloprotease motif of Jab1/Csn5 in cleavage of Nedd8 from Cul1. *Science* **298,** 608–611 (2002).
9. Fischer, E. S. *et al.* The molecular basis of CRL4$^{\mathrm{DDB2/CSA}}$ ubiquitin ligase architecture, targeting, and activation. *Cell* **147,** 1024–1039 (2011).
10. Enchev, R. I. *et al.* Structural basis for a reciprocal regulation between SCF and CSN. *Cell Rep.* **2,** 616–627 (2012).
11. Emberley, E. D., Mosadeghi, R. & Deshaies, R. J. Deconjugation of Nedd8 from Cul1 is directly regulated by Skp1-F-box and substrate, and the COP9 signalosome inhibits deneddylated SCF by a noncatalytic mechanism. *J. Biol. Chem.* **287,** 29679–29689 (2012).
12. Zimmerman, E. S., Schulman, B. A. & Zheng, N. Structural assembly of cullin-RING ubiquitin ligase complexes. *Curr. Opin. Struct. Biol.* **20,** 714–721 (2010).
13. Furukawa, M., Zhang, Y., McCarville, J., Ohta, T. & Xiong, Y. The CUL1 C-terminal sequence and ROC1 are required for efficient nuclear accumulation, NEDD8 modification, and ubiquitin ligase activity of CUL1. *Mol. Cell. Biol.* **20,** 8185–8197 (2000).
14. Podust, V. N. *et al.* A Nedd8 conjugation pathway is essential for proteolytic targeting of p27$^{\mathrm{Kip1}}$ by ubiquitination. *Proc. Natl Acad. Sci. USA* **97,** 4579–4584 (2000).
15. Read, M. A. *et al.* Nedd8 modification of Cul-1 activates SCF$^{\beta\mathrm{TrCP}}$-dependent ubiquitination of IκBα. *Mol. Cell. Biol.* **20,** 2326–2333 (2000).

16. Wu, K., Chen, A. & Pan, Z. Q. Conjugation of Nedd8 to CUL1 enhances the ability of the ROC1–CUL1 complex to promote ubiquitin polymerization. *J. Biol. Chem.* **275,** 32317–32324 (2000).
17. Morimoto, M., Nishida, T., Honda, R. & Yasuda, H. Modification of cullin-1 by ubiquitin-like protein Nedd8 enhances the activity of SCF$^{\mathrm{skp2}}$ toward p27$^{\mathrm{kip1}}$. *Biochem. Biophys. Res. Commun.* **270,** 1093–1096 (2000).
18. Sharon, M. *et al.* Symmetrical modularity of the COP9 signalosome complex suggests its multifunctionality. *Structure* **17,** 31–40 (2009).
19. Wei, N., Serino, G. & Deng, X. W. The COP9 signalosome: more than a protease. *Trends Biochem. Sci.* **33,** 592–600 (2008).
20. Zhao, R. *et al.* Subunit 6 of the COP9 signalosome promotes tumorigenesis in mice through stabilization of MDM2 and is upregulated in human cancers. *J. Clin. Invest.* **121,** 851–865 (2011).
21. Lee, M. H., Zhao, R., Phan, L. & Yeung, S. C. Roles of COP9 signalosome in cancer. *Cell Cycle* **10,** 3057–3066 (2011).
22. Echalier, A. *et al.* Insights into the regulation of the human COP9 signalosome catalytic subunit, CSN5/Jab1. *Proc. Natl Acad. Sci. USA* **110,** 1273–1278 (2013).
23. Enchev, R. I., Schreiber, A., Beuron, F. & Morris, E. P. Structural insights into the COP9 signalosome and its common architecture with the 26S proteasome lid and eIF3. *Structure* **18,** 518–527 (2010).
24. Beck, F. *et al.* Near-atomic resolution structural model of the yeast 26S proteasome. *Proc. Natl Acad. Sci. USA* **109,** 14870–14875 (2012).
25. Lander, G. C. *et al.* Complete subunit architecture of the proteasome regulatory particle. *Nature* **482,** 186–191 (2012).
26. Lasker, K. *et al.* Molecular architecture of the 26S proteasome holocomplex determined by an integrative approach. *Proc. Natl Acad. Sci. USA* **109,** 1380–1387 (2012).
27. da Fonseca, P. C., He, J. & Morris, E. P. Molecular model of the human 26S proteasome. *Mol. Cell* **46,** 54–66 (2012).
28. Matyskiela, M. E., Lander, G. C. & Martin, A. Conformational switching of the 26S proteasome enables substrate degradation. *Nature Struct. Mol. Biol.* **20,** 781–788 (2013).
29. Siridechadilok, B., Fraser, C. S., Hall, R. J., Doudna, J. A. & Nogales, E. Structural roles for human translation factor eIF3 in initiation of protein synthesis. *Science* **310,** 1513–1515 (2005).
30. Sun, C. *et al.* Functional reconstitution of human eukaryotic translation initiation factor 3 (eIF3). *Proc. Natl Acad. Sci. USA* **108,** 20473–20478 (2011).
31. Hashem, Y. *et al.* Structure of the mammalian ribosomal 43S preinitiation complex bound to the scanning factor DHX29. *Cell* **153,** 1108–1119 (2013).
32. Querol-Audi, J. *et al.* Architecture of human translation initiation factor 3. *Structure* **21,** 920–928 (2013).
33. Wei, Z. *et al.* Crystal structure of human eIF3k, the first structure of eIF3 subunits. *J. Biol. Chem.* **279,** 34983–34990 (2004).
34. Dessau, M. *et al.* The *Arabidopsis* COP9 signalosome subunit 7 is a model PCI domain protein with subdomains involved in COP9 signalosome assembly. *Plant Cell* **20,** 2815–2834 (2008).
35. Pathare, G. R. *et al.* The proteasomal subunit Rpn6 is a molecular clamp holding the core and regulatory subcomplexes together. *Proc. Natl Acad. Sci. USA* **109,** 149–154 (2012).
36. Zhang, H. *et al.* The crystal structure of the MPN domain from the COP9 signalosome subunit CSN6. *FEBS Lett.* **586,** 1147–1153 (2012).
37. Boehringer, J. *et al.* Structural and functional characterization of Rpn12 identifies residues required for Rpn10 proteasome incorporation. *Biochem. J.* **448,** 55–65 (2012).
38. Lee, J. H. *et al.* Crystal structure and versatile functional roles of the COP9 signalosome subunit 1. *Proc. Natl Acad. Sci. USA* **110,** 11845–11850 (2013).
39. Khoshnevis, S. *et al.* Structural integrity of the PCI domain of eIF3a/TIF32 is required for mRNA recruitment to the 43S pre-initiation complexes. *Nucleic Acids Res.* **42,** 4123–4139 (2014).
40. Worden, E. J., Padovani, C. & Martin, A. Structure of the Rpn11–Rpn8 dimer reveals mechanisms of substrate deubiquitination during proteasomal degradation. *Nature Struct. Mol. Biol.* **21,** 220–227 (2014).
41. Pathare, G. R. *et al.* Crystal structure of the proteasomal deubiquitylation module Rpn8-Rpn11. *Proc. Natl Acad. Sci. USA* **111,** 2984–2989 (2014).
42. Hofmann, K. & Bucher, P. The PCI domain: a common theme in three multiprotein complexes. *Trends Biochem. Sci.* **23,** 204–205 (1998).
43. Ellisdon, A. M. & Stewart, M. Structural biology of the PCI-protein fold. *BioArchitecture* **2,** 118–123 (2012).
44. Tran, H. J., Allen, M. D., Lowe, J. & Bycroft, M. Structure of the Jab1/MPN domain and its implications for proteasome function. *Biochemistry* **42,** 11460–11465 (2003).
45. Ambroggio, X. I., Rees, D. C. & Deshaies, R. J. JAMM: a metalloprotease-like zinc site in the proteasome and signalosome. *PLoS Biol.* **2,** e2 (2004).
46. Pick, E. *et al.* The minimal deneddylase core of the COP9 signalosome excludes the Csn6 MPN^{-} domain. *PLoS ONE* **7,** e43980 (2012).
47. Sato, Y. *et al.* Structural basis for specific cleavage of Lys 63-linked polyubiquitin chains. *Nature* **455,** 358–362 (2008).
48. Duda, D. M. *et al.* Structural insights into NEDD8 activation of cullin-RING ligases: conformational control of conjugation. *Cell* **134,** 995–1006 (2008).
49. Zheng, N. *et al.* Structure of the Cul1–Rbx1–Skp1–F box$^{\mathrm{Skp2}}$ SCF ubiquitin ligase complex. *Nature* **416,** 703–709 (2002).
50. Hao, B. *et al.* Structural basis of the Cks1-dependent recognition of p27$^{\mathrm{Kip1}}$ by the SCF$^{\mathrm{Skp2}}$ ubiquitin ligase. *Mol. Cell* **20,** 9–19 (2005).

Supplementary Information is available in the online version of the paper.

Acknowledgements This work was supported by the Novartis Research Foundation, the Swiss National Science Foundation (31003A_144020) and the European Research Council (MoBa-CS #260481). G.M.L. is the recipient of a Novartis (NIBR) presidential postdoctoral fellowship. Part of this work was performed at beamline X06DA and X10SA of the Swiss Light Source. We thank H. Wu for sharing the *A. thaliana* CSN1 coordinates before publication; H. Gut, J. Keusch and M. Jones for support; and C. Vonrhein and the BUSTER development group, R. Read, A. McCoy, T. Terwilliger, R. Nicholls, R. Kingston and K. Diederichs for technical advice. We are also grateful for the assistance of B. Martoglio, R. Assenberg, C. Logel and I. Bechtold in developing the PT22 CSN assays.

Author Contributions G.M.L., M.R. and N.H.T. initiated the project. G.M.L. established the purification methods for CSN and CSN4, produced most proteins, performed the gel-based assays, and obtained the crystals. G.M.L. improved the crystals with input from R.D.B. G.M.L. and R.D.B. collected the diffraction data. R.D.B. carried out the crystallographic analyses and interpreted the results. S.C. performed electron microscopy analysis. D.H. carried out protein analysis. E.S.F. developed the binding and activity assays with input from U.H.; E.S.F. performed the assays and analysed the results. N.H.T. supervised all aspects of the project and analysed the results. R.D.B. and N.H.T. wrote the manuscript with important contributions from G.M.L., M.R., E.S.F. and S.C.

Author Information The coordinates and structure factors have been deposited in the Protein Data Bank under accession codes 4D10 and 4D18 for two CSN unit cell variants, and 4D0P for CSN4. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to N.H.T. (nicolas.thoma@fmi.ch).

doi:10.1038/nature13567

Three-dimensional structure of human γ-secretase

Peilong Lu[1,2]*, Xiao-chen Bai[3]*, Dan Ma[1,2]*, Tian Xie[1,2], Chuangye Yan[2,4], Linfeng Sun[1,2], Guanghui Yang[2,4], Yanyu Zhao[1,2], Rui Zhou[1,2], Sjors H. W. Scheres[3] & Yigong Shi[1,2]

The γ-secretase complex, comprising presenilin 1 (PS1), PEN-2, APH-1 and nicastrin, is a membrane-embedded protease that controls a number of important cellular functions through substrate cleavage. Aberrant cleavage of the amyloid precursor protein (APP) results in aggregation of amyloid-β, which accumulates in the brain and consequently causes Alzheimer's disease. Here we report the three-dimensional structure of an intact human γ-secretase complex at 4.5 Å resolution, determined by cryo-electron-microscopy single-particle analysis. The γ-secretase complex comprises a horseshoe-shaped transmembrane domain, which contains 19 transmembrane segments (TMs), and a large extracellular domain (ECD) from nicastrin, which sits immediately above the hollow space formed by the TM horseshoe. Intriguingly, nicastrin ECD is structurally similar to a large family of peptidases exemplified by the glutamate carboxypeptidase PSMA. This structure serves as an important basis for understanding the functional mechanisms of the γ-secretase complex.

γ-Secretase is a membrane-embedded aspartyl protease that cleaves a large number of transmembrane substrate proteins within their membrane-spanning regions, with the cleavage products serving as signalling molecules[1,2]. This process is known as regulated intramembrane proteolysis (RIP)[3]. Two extensively studied substrates of γ-secretase are the amyloid precursor protein (APP) and the Notch receptor[2]. Successive cleavages of APP give rise to several amyloid-β peptides, each with different length[4]. Aberrant accumulation of an aggregation-prone 42-residue amyloid-β ($A\beta_{42}$) over a 40-residue product ($A\beta_{40}$) leads to formation of amyloid-β plaques in the brain, triggering the development and pathogenesis of Alzheimer's disease[2]. Cleavage of the Notch receptor results in the release and translocation of its intracellular domain into the nucleus[2]. Abnormal Notch signalling is linked to developmental defects and several types of cancer[2].

The γ-secretase complex consists of four components: PS1, PEN-2, APH-1 and nicastrin, each containing at least one predicted transmembrane segment (TM)[5,6]. Together, these proteins have a molecular weight of approximately 170 kilodaltons (kDa), whereas the nicastrin ECD has an additional 30–70 kDa of glycosylation[7]. Presenilin is the catalytic component and contains nine TMs[8–11]. Association with PEN-2 facilitates an autocatalytic cleavage of presenilin between TM6 and TM7, producing two fragments known as the amino-terminal fragment (NTF) and the carboxy-terminal fragment (CTF)[12,13]. APH-1 and nicastrin assemble into a stable subcomplex[14,15], which then interacts with the CTF of presenilin[16,17]. Nicastrin contains a large extracellular domain that is thought to be responsible for substrate recruitment[18,19]. The central role of presenilin in the γ-secretase complex is evidenced by the identification of over 150 missense mutations[2], each derived from a patient with Alzheimer's disease.

Despite advances in understanding of the functional aspects of γ-secretase, structural characterization has been extremely slow, owing mainly to the daunting challenges of expression and purification of the intact γ-secretase. The limited structural information on γ-secretase is restricted to low-resolution images derived from electron microscopy analysis[20–24], a nuclear magnetic resonance (NMR) structure of the CTF of presenilin[25], and a crystal structure of an archaeal homologue of presenilin[26]. Consequently, there is little mechanistic understanding of the γ-secretase functions.

During the past several years, we have made rigorous efforts to prepare homogeneous, active human γ-secretase for structural investigation. We attempted cryo-electron-microscopy (cryo-EM) single-particle reconstruction by exploiting technological advances in direct electron detection and statistical image processing[27,28]. Recent applications of this rapidly developing technology include near 3 Å resolution structures of a mitochondrial ribosome large subunit[29], the 12-fold symmetric F_{420}-reducing hydrogenase[30], and the fourfold symmetric TRPV1 complex[31]. Despite these advances, near-atomic resolution reconstruction remains challenging for smaller, non-symmetric proteins such as human γ-secretase. In this study, we report a three-dimensional structure of this membrane-embedded complex with an overall resolution of 4.5 Å, which reveals its domain architecture, secondary structural elements, TM arrangement, and ECD fold, and provides important functional insights.

Preparation of the γ-secretase complex

The human APH-1 is encoded by two genes, *APH-1A* and *APH-1B*, of which *APH-1A* seems to be more important[32]. Similarly, human presenilin has two forms: PS1 and PS2, and PS1 contains the vast majority of disease-derived mutations[33]. Owing to these considerations, we focused our effort on the human γ-secretase that comprises PS1, PEN-2, APH-1aL (the major form of APH-1; referred to hereafter as APH-1) and nicastrin. We initially assembled a systematic effort to examine the expression levels of the individual components, select subcomplexes, as well as the intact γ-secretase complex in four different expression systems: bacteria, yeast, insect cells and mammalian cells. We succeeded in transient co-expression of all four components of the human γ-secretase complex in mammalian HEK293F cells. The coding sequences of PS1,

[1] Ministry of Education Key Laboratory of Protein Science, Center for Structural Biology, School of Life Sciences and School of Medicine, Tsinghua University, Beijing 100084, China. [2] Tsinghua-Peking Joint Center for Life Sciences, Center for Structural Biology, School of Life Sciences and School of Medicine, Tsinghua University, Beijing 100084, China. [3] MRC Laboratory of Molecular Biology, Cambridge Biomedical Campus, Cambridge CB2 0QH, UK. [4] State Key Laboratory of Bio-membrane and Membrane Biotechnology, Center for Structural Biology, School of Life Sciences and School of Medicine, Tsinghua University, Beijing 100084, China.
*These authors contributed equally to this work.

PEN-2, APH-1 and nicastrin were individually cloned into our custom-designed pMLink plasmid, in which expression of each of the four γ-secretase components was under a separate promotor control (Extended Data Fig. 1a, b). The resulting pMLink plasmid was transfected into HEK293F cells (Fig. 1a).

To facilitate purification, we used a range of different affinity tags to label the N or C termini of the four individual components. The best outcome was achieved with a Flag tag at the N terminus of PEN-2. The γ-secretase-containing membrane fractions of HEK293F cells, extracted by the detergent CHAPSO, was purified over an anti-Flag affinity resin and further fractionated on a size exclusion column (Fig. 1a, b). The resulting γ-secretase complex exhibited excellent solution behaviour and could be easily visualized on SDS–PAGE by Coomassie blue staining, free of any major contaminating protein. Importantly, the NTF and CTF were clearly visible, suggesting completion of PS1 autoproteolysis in the presence of the other three components. By contrast, expression of PS1 alone yielded the intact, uncleaved protein (Extended Data Fig. 1c).

Presence of the NTF and CTF is indicative of an active γ-secretase complex. To examine this, we reconstituted a γ-secretase activity assay using the substrate APP-C100, which contains the C-terminal 100 amino acids of APP[34]. Incubation of γ-secretase with the substrate in a 1:10 molar ratio led to generation of APP intracellular domain (AICD) (Fig. 1c). The presenilin-specific inhibitor III-31C (ref. 35), but not DMSO (dimethylsulphoxide), blocked the cleavage of APP-C100. The level of γ-secretase activity is similar to what had been reported[18]. The same conclusion was obtained for γ-secretase in the presence of amphipol A8-35 under the same buffer condition as used in later cryo-EM analysis (Extended Data Fig. 1d). We concluded that the human γ-secretase was in an active conformation. Nevertheless, there is a possibility that, given sample manipulation, the electron-microscopy structure described below may not represent the fully active conformation.

Cryo-EM analysis of γ-secretase

Initial attempts to image γ-secretase in digitonin using an FEI Falcon-II direct-electron detector produced a three-dimensional reconstruction with a large disc-shaped 'body' and a protruding 'head', which could accommodate the TMs and extracellular domains of γ-secretase, respectively (Extended Data Fig. 2a, d). However, despite sharp contrast in the individual particles, this reconstruction showed few internal features. The TMs were not clearly resolved, and the strongest density appeared at the periphery of the disc-shaped body, which is likely to have derived from the detergent digitonin. These results concurred with relatively poor accuracies in the alignment of the particles as estimated in the employed statistical refinement procedure[36], and suggested that the disordered nature of the detergent molecules and the small size of the complex precluded correct alignment of the particles.

To minimize the effect of the disordered detergent on refinement, we replaced digitonin with amphipol A8-35 (Extended Data Fig. 2b, d). In addition, we also imaged these samples using a Gatan K2 Summit direct-electron detector in single-electron counting mode to achieve higher signal-to-noise ratios at the lower spatial frequencies, which are crucial for particle alignment. Combined with statistical image classification and movie processing[27], this approach produced a markedly improved map with an overall resolution of 4.5 Å (Fig. 2a and Extended Data Figs 2c,d and 3).

At this resolution, 19 TMs were identified, the β-strands in the nicastrin ECD were well-resolved, and side-chain densities started to show for portions of the nicastrin ECD and some of the TMs (Fig. 2a). Densities for some of the linker sequences between neighbouring TMs were improved by further image classification, which led to a map with an overall resolution of 5.4 Å from a subset of the particles (Extended Data Fig. 3b). The overall correctness of the density map and its handedness were confirmed by the tilt-pair test[37] (Extended Data Fig. 4).

Overall structure of the γ-secretase

The 19 TMs are organized into a horseshoe-shaped structure (Fig. 2b, top panel). In contrast to the density for the TMs, the density for the connecting sequences between neighbouring TMs is weak or absent, possibly reflecting the disordered nature in these hydrophilic loops. Nevertheless, at least seven TMs are connected by strong density (Extended Data Fig. 5), suggesting their order of linkage in γ-secretase. For ease of discussion, we numbered the 19 TMs (Fig. 2b, bottom panel). These TMs exhibit quite different lengths, with two connected TMs (TM17 and TM18) protruding halfway into the membrane from the cytoplasmic side (Fig. 2b and Extended Data Fig. 5). Two bent TMs (TM6 and TM7) are placed on the concave side of the horseshoe, facing the hollow centre. The large, empty pocket seems to be poised for binding to some structural element; perhaps the substrate protein.

The distribution of the 19 TMs is uneven, with considerably more TMs concentrated on one end of the horseshoe-shaped structure (referred to as the 'thick' end) than the other end (the 'thin' end). In the thin end, there are no more than two layers of TMs when viewed perpendicular to the membrane (Fig. 2b, bottom panel). By contrast, the thick end has at least three layers of TMs. The archaeal homologue of PS1, mmPSH, exhibits a relatively complex membrane topology, with three layers of TMs in an inactive conformation[26]. Assuming all TMs in the γ-secretase have been identified in the current electron-microscopy maps, this analysis suggests that PS1 might be located within the thick end of the TM horseshoe.

There is a large region of well-defined density outside the membrane-spanning region, and the density vastly exceeds that of the sequences from γ-secretase on the intracellular side. Among the four components of γ-secretase, nicastrin is the only one that has a sizable ECD, and most of the extracellular density is thus attributable to nicastrin (Fig. 2). Intriguingly, nicastrin ECD is located immediately above the hollow centre of the TM horseshoe and interacts closely with the extracellular loops of several TMs on both ends of the horseshoe (Fig. 2b). This

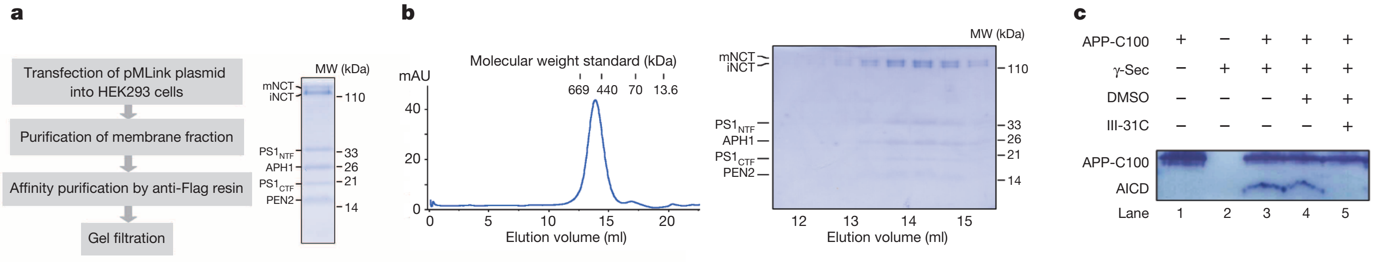


Figure 1 | Expression and purification of active human γ-secretase. **a**, A schematic diagram of the protocol for the expression and purification of the intact human γ-secretase complex. pMLink is our custom-designed vector for simultaneous co-expression of multiple proteins in mammalian cells. **b**, A representative gel-filtration chromatography of human γ-secretase. The peak fractions were visualized on SDS–PAGE by Coomassie staining. PS1 had been completely autoproteolysed into NTF and CTF, whereas nicastrin (NCT) existed in two forms: immature (iNCT) and mature (mNCT), reflecting differences in glycosylation. **c**, The purified γ-secretase was proteolytically active against the APP substrate C100. Cleavage of the substrate APP-C100 was blocked by the specific inhibitor III-31C.

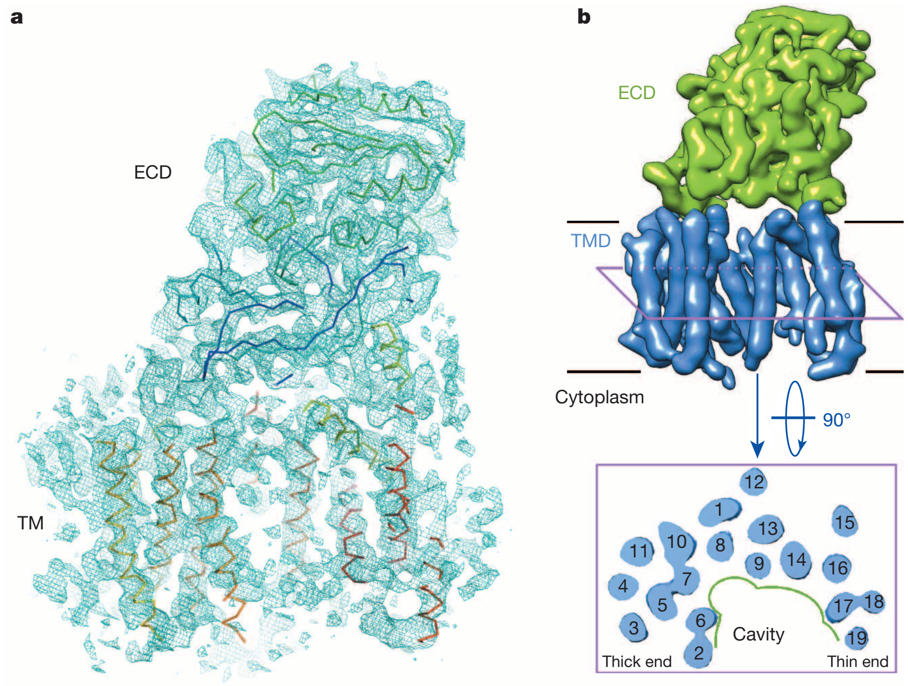


Figure 2 | Overall structure of the human γ-secretase complex. a, An overall density map for the entire human γ-secretase complex. α-Carbon traces are shown for some of the TMs and the ECD. The 5.4-Å map was used in both **a** and **b**. The electron-microscopy maps are coloured cyan. **b**, Overall structure of the human γ-secretase complex. Structure of the γ-secretase is viewed from within the plane of lipid membrane (upper panel). The 19 TMs from the four components of γ-secretase are coloured blue, whereas the ECD of nicastrin is shown in green. A cut-through section of the 19 TMs in γ-secretase is shown in the bottom panel. The TMs form a horseshoe-shaped structure, with more TMs concentrated at the thick end. The TMs are numbered arbitrarily, for ease of discussion, from 1 to 19. Figures 2a and 3 were prepared using PyMol[49], and Fig. 2b was made in Chimera[50].

organization is consistent with the reported function of substrate recruitment for nicastrin[18].

Human γ-secretase contains four full-length proteins: PS1 (residues 1–467), PEN-2 (residues 1–101), APH-1 (residues 1–265), and nicastrin (residues 1–709). With glycosylation of nicastrin, the predicted molecular weight of the intact γ-secretase is approximately 230 kDa (ref. 22). The observed density accounts for approximately half of the total molecular weight of γ-secretase, with the 19 TMs accommodating about 500 residues and nicastrin ECD containing about 650 residues. The lack of obvious density for the other sequences is likely to reflect their flexible nature, including the 30–70 kDa of oligosaccharides on glycosylated residues in the nicastrin ECD. Only 43 of the 181 residues predicted to be on the cytoplasmic side of PS1 are hydrophobic, representing 24 per cent of the total sequences and unlikely to be sufficient for formation of a stable structural core. In addition, the extracellular sequences for PEN-2 (residues 1–19 and 78–101) are predicted to be hydrophilic and flexible. These residues are missing in the current maps.

Structure of nicastrin ECD

Nicastrin ECD was previously predicted to conform to the aminopeptidase superfamily fold[38]. The relatively high-resolution features in the density for nicastrin ECD (Fig. 2a, Fig. 3a and Extended Data Fig. 6) prompted us to pursue a model for its domain architecture. To facilitate this task, we searched for sequences in the Protein Data Bank (PDB) that are homologous to those of nicastrin ECD. One of the matches was the glutamate carboxyl peptidase PSMA (PDB code 2XEF (ref. 39)) (Extended Data Fig. 7), confirming the earlier prediction[38]. Of the 218 aligned amino acids between PSMA and nicastrin, 52 are identical and 80 are similar. Visual inspection of the extracellular electron-microscopy density revealed an excellent match to the structure of PSMA[39]. The conserved topology between these two structures allowed tracing of approximately 400 residues with side chains and 20 residues as poly-Ala sequences in the nicastrin ECD (Fig. 3b). The remaining unassigned electron-microscopy density is relatively poor and can accommodate about 200 residues. The modelled part of nicastrin ECD resembles a dumbbell, with a large lobe and a small lobe (Fig. 3b), which can be superimposed with those of PSMA with root-mean-squared deviations (r.m.s.d.) of approximately 2.6 and 3.6 Å over 231 and 111 aligned α-carbon (Cα) atoms, respectively (Fig. 3c).

Discussion

The structural homology between nicastrin ECD and the peptidase PSMA may not support the possibility that nicastrin could serve as an active protease in cells. PSMA is a zinc metalloprotease and the majority of the residues that coordinate the two zinc ions in PSMA have been replaced in nicastrin (Extended Data Fig. 7). Moreover, we have been unable to detect any protease activity for nicastrin ECD *in vitro* using a variety of potential substrate proteins under diverse conditions. Nevertheless, the fact that nicastrin ECD shares a conserved fold as PSMA and other peptidases supports the idea that nicastrin may be involved in substrate recruitment[18,19]. Nicastrin ECD seems to contain a surface groove approximately 40 Å above the lipid membrane, facing the hollow centre of the TM horseshoe (Fig. 3d). This surface groove could be a putative substrate-binding site. Because the active site of PS1 is predicted to be located approximately 20 Å below the surface of the lipid membrane[26], the putative substrate-binding site is at least 60 Å away from the catalytic Asp residues in PS1. Assuming the N terminus of the BACE cleavage product APP-C99 is recognized by this surface groove[18], a distance of 60 Å can be conveniently spanned by the primary cleavage products of APP-C99; Aβ_{40} and Aβ_{42}. Supporting this analysis, Glu 333, which was thought to be responsible for substrate binding[18,19], resides at the centre of the groove (Fig. 3d). The residue in PSMA that corresponds to Glu 333 of nicastrin is directly involved in zinc binding[39].

The lack of side-chain features in the density for the 19 TMs does not allow assignment of the four components. The weak density for the loops connecting neighbouring TMs further complicates the assignment task. Nevertheless, we suggest a speculative assignment, in which all 9 TMs of PS1 are located within the thick end of the TM horseshoe

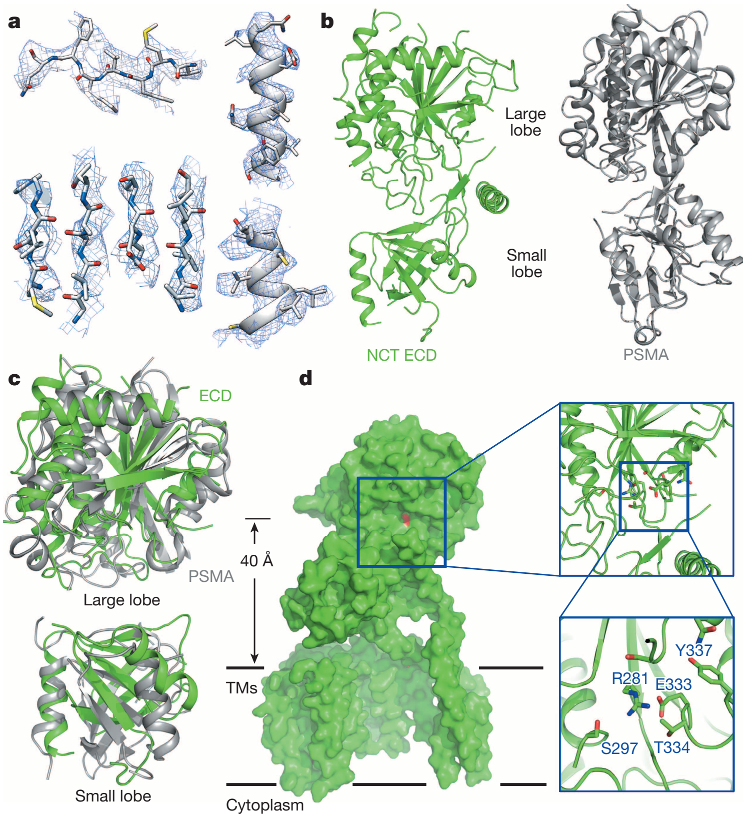


Figure 3 | Structure of the extracellular domain of nicastrin. **a**, Representative cryo-EM density for β-strands (left panels) and α-helices (right panels) of the nicastrin ECD. The 4.5-Å map was used here. **b**, The overall structure of nicastrin ECD closely resembles that of the glutamate carboxyl peptidase PSMA[39]. The atomic model of nicastrin ECD is shown in green. The structure of PSMA is shown in grey, displayed in the right panel for comparison. **c**, Structural comparison between nicastrin (green) and PSMA[39] (grey) for the large lobe (top panel) and the small lobe (bottom panel). **d**, Identification of a putative substrate-binding site in nicastrin ECD. A surface groove on nicastrin ECD, located 40 Å above the lipid membrane, faces the hollow centre of the TM horseshoe. Glu 333, which is thought to have an important role in substrate recruitment[18,19], resides in the groove.

(Extended Data Fig. 8a). The PS1 homologue mmPSH contains three layers of TMs[26]. Based on the current electron-microscopy density, the thick end is the only place in the horseshoe structure with three layers of TMs. The putative assignment of TM1 from PS1 was facilitated by the bent nature of TM1 in mmPSH[26]. PEN-2 was shown to be in close proximity of the CTF of PS1 (ref. 40) and to directly bind TM4 of PS1 (refs 41, 42), and APH-1 and nicastrin were thought to interact with the CTF of PS1 (refs 16, 17); these features are recapitulated in our model. In this speculative model, TM6 and TM7 of PS1, which harbour the catalytic Asp residues, and TM9, which contains the substrate recognition sequence, face the hollow centre of the TM horseshoe (Extended Data Fig. 8a). This analysis suggests the location of substrate cleavage by γ-secretase. The two TMs of PEN-2 are likely to be inserted between TM7 and TM8 on the cytoplasmic side, leading to a major conformational rearrangement of the TMs in PS1 compared to those in mmPSH[26] and opening of the putative substrate entry site (Extended Data Fig. 8b). This analysis might explain why PS1 autoproteolysis only occurs in the presence of PEN-2. Despite the charm of this model, we cannot rule out the opposing possibility, whereby PS1 is placed in the thin end (Extended Data Fig. 8c). After all, it remains to be seen whether mmPSH represents a sound structural model for PS1 or whether all TMs in the γ-secretase have been identified.

Although the overall resolution of our structure is 4.5 Å, the resolution for the TMs is lower and thus does not allow modelling of specific side chains. Compared to the detergent choice of digitonin, amphipol was clearly better in the cryo-EM analysis of γ-secretase and helped to improve the quality of image reconstruction. The use of amphipol was reported previously in at least two cryo-EM studies of membrane proteins[31,43]. As a new class of surfactants designed to improve the solubility of membrane proteins[44], amphipols may prove to be an important tool for future electron-microscopy-based investigation of many other membrane proteins.

Recent structural investigations of intramembrane proteases such as the prokaryotic homologues of rhomboid[45–47], S2P[48] and presenilin[26] have provided hints about the functional mechanisms of these membrane-embedded signalling proteases. In this study, we report the first cryo-EM density map of human γ-secretase in which individual β-strands are clearly separated in nicastrin ECD and 19 TMs form a horseshoe-like structure. Our observed structural features are different from those that were derived from previous low-resolution electron-microscopy studies of γ-secretase[20–24]. Our structure marks an important step towards elucidating the molecular mechanisms of this key enzyme whose aberrant activity engenders Alzheimer's disease.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 17 April; accepted 9 June 2014.
Published online 29 June 2014.

1. Selkoe, D. J. & Wolfe, M. S. Presenilin: running with scissors in the membrane. *Cell* **131,** 215–221 (2007).
2. De Strooper, B., Iwatsubo, T. & Wolfe, M. S. Presenilins and γ-secretase: structure, function, and role in Alzheimer disease. *Cold Spring Harb. Persp. Medi.* **2,** a006304 (2012).
3. Brown, M. S., Ye, J., Rawson, R. B. & Goldstein, J. L. Regulated intramembrane proteolysis: a control mechanism conserved from bacteria to humans. *Cell* **100,** 391–398 (2000).
4. Wolfe, M. S. Toward the structure of presenilin/γ-secretase and presenilin homologs. *Biochim. Biophys. Acta* **1828,** 2886–2897 (2013).

5. De Strooper, B. Aph-1, Pen-2, and Nicastrin with Presenilin generate an active γ-secretase complex. *Neuron* **38,** 9–12 (2003).
6. Kimberly, W. T. *et al.* γ-secretase is a membrane protein complex comprised of presenilin, nicastrin, Aph-1, and Pen-2. *Proc. Natl Acad. Sci. USA* **100,** 6382–6387 (2003).
7. Schedin-Weiss, S., Winblad, B. & Tjernberg, L. O. The role of protein glycosylation in Alzheimer disease. *FEBS J.* **281,** 46–62 (2014).
8. Wolfe, M. S. *et al.* Two transmembrane aspartates in presenilin-1 required for presenilin endoproteolysis and γ-secretase activity. *Nature* **398,** 513–517 (1999).
9. De Strooper, B. *et al.* Deficiency of presenilin-1 inhibits the normal cleavage of amyloid precursor protein. *Nature* **391,** 387–390 (1998).
10. De Strooper, B. *et al.* A presenilin-1-dependent γ-secretase-like protease mediates release of Notch intracellular domain. *Nature* **398,** 518–522 (1999).
11. Struhl, G. & Greenwald, I. Presenilin is required for activity and nuclear access of Notch in *Drosophila*. *Nature* **398,** 522–525 (1999).
12. Thinakaran, G. *et al.* Endoproteolysis of presenilin 1 and accumulation of processed derivatives *in vivo*. *Neuron* **17,** 181–190 (1996).
13. Takasugi, N. *et al.* The role of presenilin cofactors in the γ-secretase complex. *Nature* **422,** 438–441 (2003).
14. Gu, Y. *et al.* APH-1 interacts with mature and immature forms of presenilins and nicastrin and may play a role in maturation of presenilin·nicastrin complexes. *J. Biol. Chem.* **278,** 7374–7380 (2003).
15. LaVoie, M. J. *et al.* Assembly of the γ-secretase complex involves early formation of an intermediate subcomplex of Aph-1 and nicastrin. *J. Biol. Chem.* **278,** 37213–37222 (2003).
16. Steiner, H., Winkler, E. & Haass, C. Chemical cross-linking provides a model of the γ-secretase complex subunit architecture and evidence for close proximity of the C-terminal fragment of presenilin with APH-1. *J. Biol. Chem.* **283,** 34677–34686 (2008).
17. Kaether, C. *et al.* The presenilin C-terminus is required for ER-retention, nicastrin-binding and γ-secretase activity. *EMBO J.* **23,** 4738–4748 (2004).
18. Shah, S. *et al.* Nicastrin functions as a γ-secretase-substrate receptor. *Cell* **122,** 435–447 (2005).
19. Dries, D. R. *et al.* Glu-333 of nicastrin directly participates in γ-secretase activity. *J. Biol. Chem.* **284,** 29714–29724 (2009).
20. Lazarov, V. K. *et al.* Electron microscopic structure of purified, active γ-secretase reveals an aqueous intramembrane chamber and two pores. *Proc. Natl Acad. Sci. USA* **103,** 6889–6894 (2006).
21. Ogura, T. *et al.* Three-dimensional structure of the γ-secretase complex. *Biochem. Biophys. Res. Commun.* **343,** 525–534 (2006); corrigendum **345,** 543 (2006).
22. Osenkowski, P. *et al.* Cryoelectron microscopy structure of purified γ-secretase at 12 Å resolution. *J. Mol. Biol.* **385,** 642–652 (2009).
23. Renzi, F. *et al.* Structure of γ-secretase and its trimeric pre-activation intermediate by single-particle electron microscopy. *J. Biol. Chem.* **286,** 21440–21449 (2011).
24. Li, Y. *et al.* Structural interactions between inhibitor and substrate docking sites give insight into mechanisms of human PS1 complexes. *Structure* **22,** 125–135 (2014).
25. Sobhanifar, S. *et al.* Structural investigation of the C-terminal catalytic fragment of presenilin 1. *Proc. Natl Acad. Sci. USA* **107,** 9644–9649 (2010).
26. Li, X. *et al.* Structure of a presenilin family intramembrane aspartate protease. *Nature* **493,** 56–61 (2013).
27. Bai, X. C., Fernandez, I. S., McMullan, G. & Scheres, S. H. Ribosome structures to near-atomic resolution from thirty thousand cryo-EM particles. *eLife* **2,** e00461 (2013).
28. Li, X. *et al.* Electron counting and beam-induced motion correction enable near-atomic-resolution single-particle cryo-EM. *Nature Methods* **10,** 584–590 (2013).
29. Amunts, A. *et al.* Structure of the yeast mitochondrial large ribosomal subunit. *Science* **343,** 1485–1489 (2014).
30. Allegretti, M., Mills, D. J., McMullan, G., Kuhlbrandt, W. & Vonck, J. Atomic model of the F420-reducing [NiFe] hydrogenase by electron cryo-microscopy using a direct electron detector. *eLife* **3,** e01963 (2014).
31. Liao, M., Cao, E., Julius, D. & Cheng, Y. Structure of the TRPV1 ion channel determined by electron cryo-microscopy. *Nature* **504,** 107–112 (2013).
32. Serneels, L. *et al.* Differential contribution of the three Aph1 genes to γ-secretase activity *in vivo*. *Proc. Natl Acad. Sci. USA* **102,** 1719–1724 (2005).
33. Tang, Y. P. & Gershon, E. S. Genetic studies in Alzheimer's disease. *Dialogues Clin. Neurosci.* **5,** 17–26 (2003).
34. Li, Y. M. *et al.* Presenilin 1 is linked with γ-secretase activity in the detergent solubilized state. *Proc. Natl Acad. Sci. USA* **97,** 6138–6143 (2000).
35. Kornilova, A. Y., Das, C. & Wolfe, M. S. Differential effects of inhibitors on the γ-secretase complex. Mechanistic implications. *J. Biol. Chem.* **278,** 16470–16473 (2003).
36. Scheres, S. H. RELION: implementation of a Bayesian approach to cryo-EM structure determination. *J. Struct. Biol.* **180,** 519–530 (2012).
37. Rosenthal, P. B. & Henderson, R. Optimal determination of particle orientation, absolute hand, and contrast loss in single-particle electron cryomicroscopy. *J. Mol. Biol.* **333,** 721–745 (2003).
38. Fagan, R., Swindells, M., Overington, J. & Weir, M. Nicastrin, a presenilin-interacting protein, contains an aminopeptidase/transferrin receptor superfamily domain. *Trends Biochem. Sci.* **26,** 213–214 (2001).
39. Zhang, A. X. *et al.* A remote arene-binding site on prostate specific membrane antigen revealed by antibody-recruiting small molecules. *J. Am. Chem. Soc.* **132,** 12711–12716 (2010).
40. Bammens, L., Chavez-Gutierrez, L., Tolia, A., Zwijsen, A. & De Strooper, B. Functional and topological analysis of Pen-2, the fourth subunit of the γ-secretase complex. *J. Biol. Chem.* **286,** 12271–12282 (2011).
41. Watanabe, N. *et al.* Pen-2 is incorporated into the γ-secretase complex through binding to transmembrane domain 4 of presenilin 1. *J. Biol. Chem.* **280,** 41967–41975 (2005).
42. Kim, S. H. & Sisodia, S. S. Evidence that the "NF" motif in transmembrane domain 4 of presenilin 1 is critical for binding with PEN-2. *J. Biol. Chem.* **280,** 41953–41966 (2005).
43. Althoff, T., Mills, D. J., Popot, J. L. & Kuhlbrandt, W. Arrangement of electron transport chain components in bovine mitochondrial supercomplex I1III2IV1. *EMBO J.* **30,** 4652–4664 (2011).
44. Tribet, C., Audebert, R. & Popot, J. L. Amphipols: polymers that keep membrane proteins soluble in aqueous solutions. *Proc. Natl Acad. Sci. USA* **93,** 15047–15050 (1996).
45. Wang, Y., Zhang, Y. & Ha, Y. Crystal structure of a rhomboid family intramembrane protease. *Nature* **444,** 179–180 (2006). Published online 2006 Oct 11.
46. Wu, Z. *et al.* Structural analysis of a rhomboid family intramembrane protease reveals a gating mechanism for substrate entry. *Nature Struct. Mol. Biol.* **13,** 1084–1091 (2006).
47. Ben-Shem, A., Fass, D. & Bibi, E. Structural basis for intramembrane proteolysis by rhomboid serine proteases. *Proc. Natl Acad. Sci. USA* **104,** 462–466 (2007).
48. Feng, L. *et al.* Structure of a site-2 protease family intramembrane metalloprotease. *Science* **318,** 1608–1612 (2007).
49. DeLano, W. L. The PyMOL Molecular Graphics System http://www.pymol.org (2002).
50. Pettersen, E. F. *et al.* UCSF Chimera—a visualization system for exploratory research and analysis. *J. Comput. Chem.* **25,** 1605–1612 (2004).

Acknowledgements We thank M. Wolfe and D. Selkoe for support and encouragement, and S. Chen, C. Savva, J. Grimmett and T. Darling for technical support. This work was funded by the Ministry of Science and Technology of China (2009CB918801 to Y.S.), National Natural Science Foundation of China (projects 30888001, 31021002 and 31130002 to Y.S.), a European Union Marie Curie Fellowship (to X.C.B.), and the UK Medical Research Council (MC_UP_A025_1013, to S.H.W.S.).

Author Contributions P.L., X.B., D.M., S.H.W.S. and Y.S. designed all experiments. P.L., X.B., D.M., T.X., C.Y., L.S., G.Y., Y.Z. and R.Z. performed the experiments. All authors contributed to data analysis. P.L., X.B., D.M., S.H.W.S. and Y.S. contributed to manuscript preparation.

Author Information The modelled atomic coordinates of nicastrin has been deposited in the Protein Data Bank with the accession code 4UPC. In addition, the 4.5 Å and 5.4 Å EM maps have been deposited in EMDB with accession codes EMD-2677 and EMD-2678, respectively. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to S. Scheres (scheres@mrc-lmb.cam.ac.uk) or Y. Shi (shi-lab@tsinghua.edu.cn).

doi:10.1038/nature13525

The origin of the local 1/4-keV X-ray flux in both charge exchange and a hot bubble

M. Galeazzi[1], M. Chiao[2], M. R. Collier[2], T. Cravens[3], D. Koutroumpa[4], K. D. Kuntz[5], R. Lallement[6], S. T. Lepri[7], D. McCammon[8], K. Morgan[8], F. S. Porter[2], I. P. Robertson[3], S. L. Snowden[2], N. E. Thomas[2], Y. Uprety[1], E. Ursino[1] & B. M. Walsh[2]†

The solar neighbourhood is the closest and most easily studied sample of the Galactic interstellar medium, an understanding of which is essential for models of star formation and galaxy evolution. Observations of an unexpectedly intense diffuse flux of easily absorbed 1/4-kiloelectronvolt X-rays[1,2], coupled with the discovery that interstellar space within about a hundred parsecs of the Sun is almost completely devoid of cool absorbing gas[3], led to a picture of a 'local cavity' filled with X-ray-emitting hot gas, dubbed the local hot bubble[4–6]. This model was recently challenged by suggestions that the emission could instead be readily produced within the Solar System by heavy solar-wind ions exchanging electrons with neutral H and He in interplanetary space[7–11], potentially removing the major piece of evidence for the local existence of million-degree gas within the Galactic disk[12–15]. Here we report observations showing that the total solar-wind charge-exchange contribution is approximately 40 per cent of the 1/4-keV flux in the Galactic plane. The fact that the measured flux is not dominated by charge exchange supports the notion of a million-degree hot bubble extending about a hundred parsecs from the Sun.

When the highly ionized solar wind interacts with neutral gas, an electron may hop from a neutral to an outer orbital of an ion, in what is known as charge exchange. The electron then cascades to the ground state of the ion, often emitting soft X-rays in the process[16]. The calculations of X-ray intensity from solar-wind charge exchange depend on limited information about heavy ion fluxes and even more uncertain atomic cross-sections. The 'Diffuse X-rays from the Local galaxy' (DXL) sounding rocket mission[17] was launched from the White Sands Missile Range in New Mexico, USA, on 12 December 2012 to make an empirical measurement of the charge exchange flux by observing a region of higher interplanetary neutral density (with a correspondingly higher charge exchange rate) called the 'helium focusing cone'. Neutral interstellar gas flows at about 25 km s^{-1} through the Solar System owing to the motion of the Sun through a small interstellar cloud. This material, mostly hydrogen atoms but about 15% helium, flows from the Galactic direction (longitude l, latitude b) ≈ (3°, 16°), placing Earth downstream of the Sun in early December[18]. The trajectories of the neutral interstellar helium atoms are governed primarily by gravity, executing hyperbolic Keplerian orbits and forming a relatively high-density focusing cone downstream of the Sun about 6° below the ecliptic plane (Fig. 1)[19]. Interstellar hydrogen, on the other hand, is also strongly affected by radiation pressure and photoionization: radiation pressure balances gravity, reducing the focusing effect, while photoionization creates a neutral hydrogen cavity around the Sun.

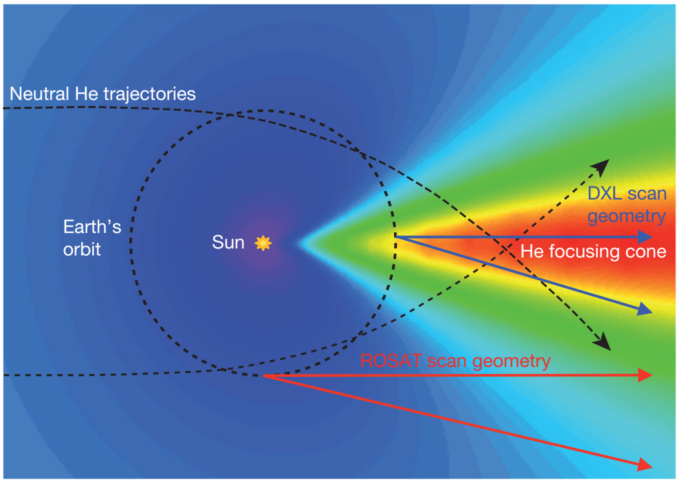


Figure 1 | The He focusing cone. Modelled interstellar He density (blue is low density; red is high density) showing the He focusing cone. Keplerian He orbits, Earth's orbit, and the DXL and ROSAT observing geometries are also shown.

The early December launch of DXL placed the He focusing cone near the zenith at midnight. The 7° field of view was scanned slowly back and forth across one side of the cone and more rapidly in a full circle to test the consistency of the derived charge exchange contribution in other directions and to make measurements of the detector particle background while DXL was looking towards Earth (Extended Data Fig. 1). Figure 2 shows the ROSAT All Sky Survey 1/4-keV map[20] (R12 band) with the paths of the DXL slow scan (red) and fast scan (white) overplotted. The ROSAT observation of the slow-scan region was performed

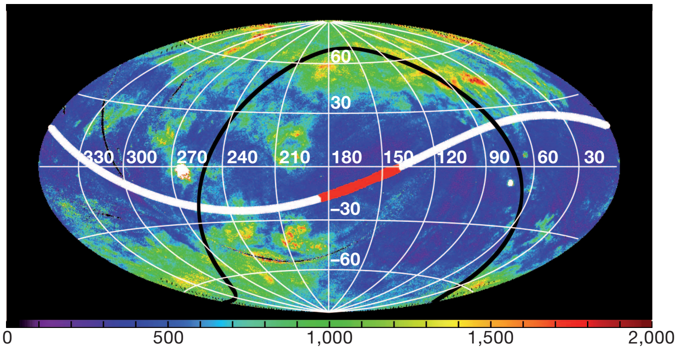


Figure 2 | The DXL scan path. ROSAT all-sky survey map in the 1/4-keV (R12) energy band, shown in Galactic coordinates (contours are labelled in degrees) with $l = 180°$, $b = 0°$ at the centre. The colour scale shows flux intensity. The units are ROSAT units, RU. The DXL scan path is the white band with the slow portion shown in red. The black line is the 90° horizon for the DXL flight. The width of the band represents the half-power diameter of the instrument beam.

[1] Department of Physics, University of Miami, Coral Gables, Florida 33124, USA. [2] NASA Goddard Space Flight Center, Greenbelt, Maryland 20771, USA. [3] Department of Physics and Astronomy, University of Kansas, Lawrence, Kansas 66045, USA. [4] Université Versailles St Quentin; Sorbonne Universités, UPMC Université Paris 06; CNRS/INSU, LATMOS-IPSL, Guyancourt 78280, France. [5] The Henry A. Rowland Department of Physics and Astronomy, Johns Hopkins University, Baltimore, Maryland 21218, USA. [6] GEPI Observatoire de Paris, CNRS UMR 8111, Université Paris Diderot, 92190, Meudon, France. [7] Department of Atmospheric, Oceanic and Space Sciences, University of Michigan, Ann Arbor, Michigan 48109, USA. [8] Department of Physics, University of Wisconsin, Madison, Wisconsin 53706, USA.
†Present address: Space Sciences Laboratory, University of California, Berkeley, California 94720, USA.

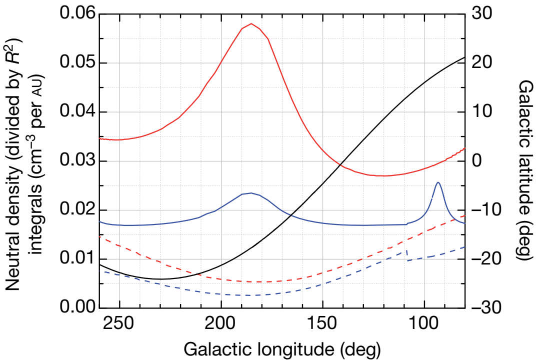


Figure 3 | Neutral atom column density for DXL and ROSAT. Neutral column density distribution integrals for each line of sight along the scan path. The density distribution in the integrals is weighted by one over the distance from the Sun squared ($1/R^2$) to reflect the dilution of the solar wind as it flows outward. The red lines are the integral for He (solid) and H (dashed) in the DXL geometry. The blue lines represent the integral for He (solid) and H (dashed) in the ROSAT geometry. The black line shows the Galactic latitude during the scan. DXL is significantly more affected by the He focusing cone, while in both cases the H contribution is small.

Table 1 | Best-fit model parameters

ROSAT geocoronal solar-wind charge exchange (RU)	0	50	0	50
α_H/α_{He}	1	1	2	2
Counter-I	0.91 ± 0.06	0.96 ± 0.06	0.87 ± 0.07	0.92 ± 0.07
Counter-II	0.97 ± 0.07	1.02 ± 0.07	0.92 ± 0.07	0.99 ± 0.07
$n_p(R_o)v_{rel}\alpha_{He}$ (RU cm^3 AU)	5,223 ± 770	3197 ± 720	4,500 ± 490	3180 ± 460
DXL/ROSAT solar-wind flux	0.63 ± 0.13	0.91 ± 0.25	0.73 ± 0.14	0.96 ± 0.20
χ^2 (136 degrees of freedom)	207	196	206	194
Total solar-wind charge exchange contribution to ROSAT R12 data	(39 ± 6)%	(37 ± 8)%	(39 ± 4)%	(42 ± 6)%

Summary of best-fit parameters for different assumptions for the geocoronal contribution to the ROSAT R12 band, and the ratio between hydrogen and helium compound cross-sections α_H/α_{He}. Counter-I and Counter-II are the ratios of the fitted DXL response to the nominal value from laboratory calibrations (the corrections are well within the range expected from spectral uncertainties). v_{rel} is the relative speed between solar wind and neutral flow and n_p is the proton density. For the ratio of solar-wind fluxes during the DXL and ROSAT measurements, we note that although both missions were near the solar maximum (and therefore should have similar isotropic composition), solar activity in terms of sunspots was weaker during the DXL measurement, as reflected by the fitted ratios. The total solar-wind charge exchange contribution to ROSAT (interplanetary + geocoronal) is divided by the observed R12 rate and given as a percentage of R12 at $b = 0°$ (333 RU, which is the lowest anywhere on the scan and close to the lowest on sky). Errors are 1σ.

in September 1990 when the line of sight was about one astronomical unit (the Earth–Sun distance, 1 AU) away from, and parallel to, the He cone, so its charge exchange contribution was not strongly affected by the cone enhancement (Fig. 1).

For this analysis, we chose pulse height limits for both of the DXL proportional counters (Counter-I and Counter-II) to match the pulse heights of the ROSAT 1/4-keV band as closely as possible (Extended Data Fig. 2). This energy range is dominated by and contains most of the emission from solar-wind charge exchange and/or the local hot bubble. To quantify the solar-wind charge exchange emission we compared both DXL and ROSAT count rates to well determined models of the interplanetary neutral distribution along the lines of sight for both sets of measurements (Fig. 3)[17,21]. Figure 4 shows the DXL and ROSAT count rates along the DXL scan path as functions of Galactic longitude. The figure shows the combined Counter-I and Counter-II count rates (black dots) during the DXL scan and the ROSAT 1/4-keV count rates in the same directions (blue solid line). The best fit to the DXL total count rate (red solid line), and the solar-wind charge exchange contributions to DXL (red dashed line) and ROSAT (blue dashed line) rates are also shown (see Table 1 for best-fit parameters: the model shown corresponds to the second column). There is potentially an additional contribution from charge exchange between the solar-wind ions and the geocoronal hydrogen surrounding Earth, which tracks the short-term variations in solar-wind flux. Time variations of a few days or less were removed from the ROSAT maps, and the current best estimate of the residual from geocoronal charge exchange is about 50 ROSAT units (1 RU = 10^{-6} counts s^{-1} arcmin^{-2}) for the ROSAT 1/4-keV band (K.D.K., J. Carter, M.P.C., Y. M. Colladovega, M.R.C., T.E.C., D.K., F.S.P., A. Read, I.P.R., D. G. Sibeck, S. F. Sembay, S.L.S., N.E.T. and D.M.W., manuscript in preparation). The geocoronal contribution to the DXL flux should be negligible, owing to the look direction, which is directly away from the Sun. The signature of the cone enhancement in the DXL data compared to the ROSAT rates is evident, highlighting the contribution from charge exchange. However, the best fit shows that the total charge exchange contribution to ROSAT is only about 40% ± 5% (statistical error) ± 5% (systematic error) of the total flux observed at the Galactic plane. Its contribution to the ROSAT flux over the DXL scan path is typically about 140 RU. For comparison, the total ROSAT 1/4-keV flux ranges from around 300 RU to 400 RU in the plane up to 1,400 RU in the brightest areas at intermediate and high latitudes. This result implies that the measured fluxes are dominated by interstellar emission, strengthening the original idea of a hot bubble filling the local interstellar medium for a hundred parsecs or so in all directions from the Sun.

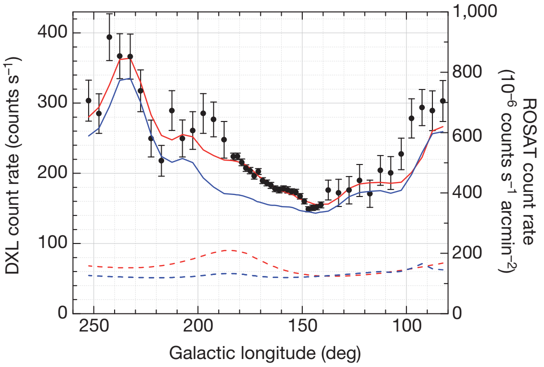


Figure 4 | Fit to DXL and ROSAT data. Combined Counter-I and Counter-II count rates (black dots) during the DXL scan and ROSAT 1/4-keV count rate in the same directions (blue solid line). The best fit to the DXL total count rate (red solid line), and the solar-wind charge exchange contribution to DXL (red dashed line) and ROSAT 1/4-keV bands (blue dashed line) are also shown. The error bars are s.e.m.

It has been pointed out that a hot bubble creates an apparent pressure balance problem with the tenuous warm cloud that the Sun is passing through. However, recent results on the magnetic contribution to the cloud pressure[22] and new three-dimensional maps of the local interstellar medium[23] bring the implied pressure of the plasma in the local hot bubble to rough agreement with pressures derived for the local interstellar clouds when the measured contribution from the solar-wind charge exchange is removed from the local hot bubble emission[24].

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 20 March; accepted 14 May 2014.

Published online 27 July 2014.

1. Bowyer, C. S., Field, G. B. & Mack, J. E. Detection of an anisotropic soft X-ray background flux. *Nature* **217,** 32–34 (1968).
2. Bunner, A. N. *et al.* Soft X-ray background flux. *Nature* **223,** 1222–1226 (1969).
3. Jenkins, E. B. & Meloy, D. A. A survey with Copernicus of interstellar O VI absorption. *Astrophys. J.* **193,** L121–L125 (1974).
4. Sanders, W. T., Kraushaar, W. L., Nousek, J. A. & Fried, P. M. Soft diffuse X rays in the southern Galactic hemisphere. *Astrophys. J.* **217,** L87–L91 (1977).
5. Cox, D. P. & Anderson, P. A. Extended adiabatic blast waves and a model of the soft X-ray background. *Astrophys. J.* **253,** 268–289 (1982).

6. Snowden, S. L., Cox, D. P., McCammon, D. & Sanders, W. T. A model for the distribution of material generating the soft X-ray background. *Astrophys. J.* **354,** 211–219 (1990).
7. Cox, D. P. Modeling the local bubble. *Lecture Notes Phys.* **506,** 121–131 (1998).
8. Cravens, T. E. Heliospheric X-ray emission associated with charge transfer of the solar wind with interstellar neutrals. *Astrophys. J.* **532,** L153–L156 (2000).
9. Lallement, R. The heliospheric soft X-ray emission pattern during the ROSAT survey: inferences on local bubble hot gas. *Astron. Astrophys.* **418,** 143–150 (2004).
10. Welsh, B. Y. & Lallement, R. Highly ionized gas in the local ISM: some like it hot? *Astron. Astrophys.* **436,** 615–632 (2005).
11. Koutroumpa, D., Lallement, R., Raymond, J. C. & Kharchenko, V. The solar wind charge-transfer X-Ray emission in the 1/4 keV energy range: inferences on local bubble hot gas at low z. *Astrophys. J.* **696,** 1517–1525 (2009).
12. Frisch, P. C. The nearby interstellar medium. *Nature* **293,** 377–379 (1981).
13. Cox, D. P. & Snowden, S. Perspective on the local interstellar medium. *Adv. Space Res.* **6,** 97–107 (1986).
14. Lallement, R. The local interstellar medium: peculiar or not? *Space Sci. Rev.* **130,** 341–353 (2007).
15. Welsh, B. Y. & Shelton, R. L. The trouble with the local bubble. *Astrophys. Space Sci.* **323,** 1–16 (2009).
16. Cravens, T. E. Comet Hyakutake X-ray source: charge transfer of solar wind heavy ions. *Geophys. Res. Lett.* **24,** 105–108 (1997).
17. Galeazzi, M. *et al.* DXL: a sounding rocket mission for the study of solar wind charge exchange and local hot bubble X-ray emission. *Exp. Astron.* **32,** 83–99 (2011).
18. Möbius, E. *et al.* Synopsis of the interstellar He parameters from combined neutral gas, pickup ion and UV scattering observations and related consequences. *Astron. Astrophys.* **426,** 897–907 (2004).
19. Frisch, P. C. The galactic environment of the Sun. *J. Geophys. Res.* **105,** 10279–10290 (2000).
20. Snowden, S. L. *et al.* First maps of the soft X-ray diffuse background from the ROSAT XRT/PSPC All-Sky Survey. *Astrophys. J.* **454,** 643–653 (1995).
21. Koutroumpa, D. *et al.* Charge-transfer induced EUV and soft X-ray emissions in the heliosphere. *Astron. Astrophys.* **460,** 289–300 (2006).
22. Burlaga, L. F. & Ness, N. F. Voyager 1 observations of the interstellar magnetic field and the transition from the heliosheath. *Astrophys. J.* **784,** 146 (2014).
23. Puspitarini, L., Lallement, R., Vergely, J. L. & Snowden, S. L. Local ISM 3D distribution and soft X-ray background: inferences on nearby hot gas and the North Polar Spur. Preprint at http://arxiv.org/abs/1401.6899 (2014).
24. Snowden, S. L, *et al.* Pressure equilibrium between the local interstellar clouds and the local hot bubble. *Astrophys. J. Lett.*. (in the press).

Acknowledgements We thank the personnel at NASA's Wallops Flight Facility and the White Sands Military Range for their support of payload development, integration and launch, and the technical personnel at the University of Miami, NASA's Goddard Space Flight Center and the University of Michigan for their support of the instrument's development. This work was supported by NASA award numbers NNX11AF04G and NNX09AF09G. D.K. and R.L. acknowledge financial support for their activity through the programme 'Soleil Héliosphère Magnétosphère' of the French space agency CNES, and the National Program 'Physique Chimie du Milieu Interstellaire' of the Institut National des Sciences de l'Univers (INSU). M.C. and N.E.T. are employed through the Center for Research and Exploration in Space Science and Technology (CRESST) and the University of Maryland, Baltimore County, Baltimore, Maryland, USA.

Author Contributions Y.U., N.E.T., M.G., D.M., M.R.C., F.S.P. and S.T.L. contributed to hardware development. Y.U., N.E.T., M.G., D.M., M.R.C., F.S.P., M.C., D.K., K.D.K. and K.M. contributed to launch operations. M.G., D.M., Y.U., N.E.T., M.R.C., D.K., K.D.K., K.M., T.C., I.P.R., S.L.S., E.U. and B.M.W. contributed to data reduction and analysis. D.K. and R.L. prepared the neutral integral distributions. All authors discussed the results and commented on the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to M.G. (galeazzi@physics.miami.edu).

doi:10.1038/nature13632

Cohesive forces prevent the rotational breakup of rubble-pile asteroid (29075) 1950 DA

Ben Rozitis[1], Eric MacLennan[1] & Joshua P. Emery[1]

Space missions[1] and ground-based observations[2] have shown that some asteroids are loose collections of rubble rather than solid bodies. The physical behaviour of such 'rubble-pile' asteroids has been traditionally described using only gravitational and frictional forces within a granular material[3]. Cohesive forces in the form of small van der Waals forces between constituent grains have recently been predicted to be important for small rubble piles (ten kilometres across or less), and could potentially explain fast rotation rates in the small-asteroid population[4–6]. The strongest evidence so far has come from an analysis of the rotational breakup of the main-belt comet P/2013 R3 (ref. 7), although that was indirect and poorly constrained by observations. Here we report that the kilometre-sized asteroid (29075) 1950 DA (ref. 8) is a rubble pile that is rotating faster than is allowed by gravity and friction. We find that cohesive forces are required to prevent surface mass shedding and structural failure, and that the strengths of the forces are comparable to, though somewhat less than, the forces found between the grains of lunar regolith.

It is possible to infer the existence of cohesive forces within an asteroid by determining whether it is a rubble pile with insufficient self-gravity to prevent rotational breakup by centrifugal forces. One of the largest known candidates is the near-Earth asteroid (29075) 1950 DA (mean diameter of 1.3 km; ref. 8), because it has a rotation period of 2.1216 h that is just beyond the critical spin limit of about 2.2 h estimated for a cohesionless asteroid[9]. A rubble-pile structure and the degree of self-gravity can be determined by a bulk density measurement, which can be acquired through model-to-measurement comparisons of Yarkovsky orbital drift[10]. This drift arises on a rotating asteroid with non-zero thermal inertia, and is caused by the delayed thermal emission of absorbed sunlight, which applies a small propulsion force to the asteroid's afternoon side. Thermal-infrared observations can constrain the thermal inertia value[11], and precise astrometric position measurements conducted over several years can constrain the degree of Yarkovsky orbital drift[2]. Recently, the orbital semimajor axis of (29075) 1950 DA has been observed to be decreasing at a rate of 44.1 ± 8.5 m yr^{-1} because of the Yarkovsky effect[12], which indicates that the asteroid's sense of rotation must be retrograde. Using the Advanced Thermophysical Model[13,14], in combination with the retrograde radar shape model[8], archival WISE thermal-infrared data[15] (Extended Data Table 1, and Extended Data Figs 1 and 2), and orbital state[12], we determined the thermal inertia and bulk density of (29075) 1950 DA (see Methods). The thermal inertia value was found to be remarkably low at $24^{+20}_{-14}\ \mathrm{J\,m^{-2}\,K^{-1}\,s^{-1/2}}$, which gives a corresponding bulk density of 1.7 ± 0.7 g cm^{-3} (Fig. 1 and Extended Data Fig. 3). This bulk density is much lower than the minimum value of 3.5 g cm^{-3} required to prevent loss of surface material by centrifugal forces (Fig. 2).

Spectral observations of (29075) 1950 DA indicate either an E- or M-type classification in the Tholen taxonomic system[16]. However, its low optical albedo and low radar circular polarization ratio[8] rule out the E-type classification (Extended Data Table 2). The derived bulk density is inconsistent with the traditional view that M-type asteroids are metallic bodies. However, the Rosetta spacecraft encounter with main-belt asteroid (21) Lutetia has demonstrated that not all M-type asteroids are metal-rich[17]. Indeed, the low radar albedo[8] of (29075) 1950 DA is very similar to that of (21) Lutetia, suggesting a similar composition. The best meteorite analogue for (21) Lutetia is an enstatite chondrite[17], which has a grain density of 3.55 g cm^{-3}. Taking the same meteorite analogue and grain density for (29075) 1950 DA implies a macro-porosity of 51 ± 19% and indicates that it is a rubble-pile asteroid (Fig. 1).

Given that the WISE observations were taken when (29075) 1950 DA was about 1.7 AU (one astronomical unit, AU, is the distance from Earth to the Sun) from the Sun, the derived thermal inertia value scales to $36^{+30}_{-20}\ \mathrm{J\,m^{-2}\,K^{-1}\,s^{-1/2}}$ at 1 AU because of temperature-dependent effects. This scaled value is comparable to that of the ~45 $\mathrm{J\,m^{-2}\,K^{-1}\,s^{-1/2}}$ value determined for the lunar surface from thermal-infrared measurements[18], and implies the presence of a similar fine-grained regolith. This is consistent with (29075) 1950 DA's low radar circular polarization ratio, which suggests a very smooth surface at centimetre to decimetre scales[8]. The sub-observer latitude of the WISE observations was ~2°, which indicates that this surface material was primarily detected around (29075) 1950 DA's equator.

For the derived bulk density, (29075) 1950 DA has 48 ± 24% of its surface experiencing negative ambient gravity (that is, surface elements

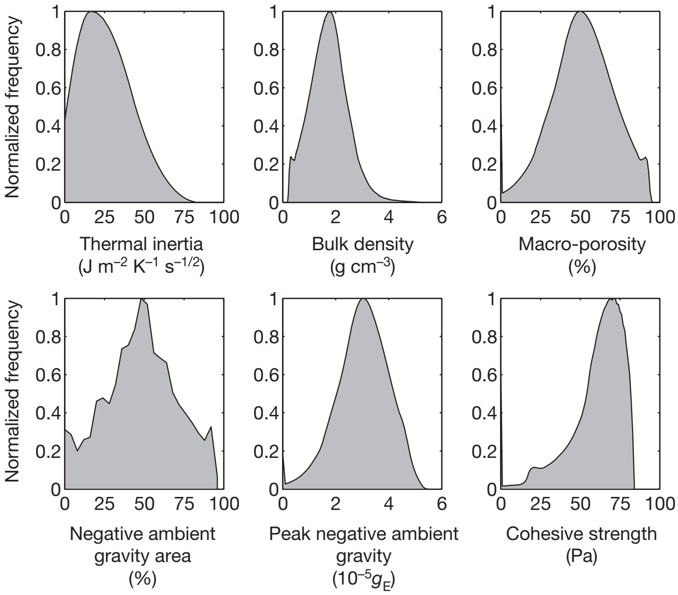


Figure 1 | Physical property distributions of (29075) 1950 DA. These were derived by the Advanced Thermophysical Model (ATPM) at the 3σ confidence level by χ^2 fitting to the 16 WISE thermal-infrared observations and to the observed rate of Yarkovsky orbital drift. The best model fit had a reduced-χ^2 value of 1.06 with a corresponding P value of 0.39. The distributions have median values and 1σ ranges of $24^{+20}_{-14}\ \mathrm{J\,m^{-2}\,K^{-1}\,s^{-1/2}}$, 1.7 ± 0.7 g cm^{-3}, 51 ± 19%, 48 ± 24%, $(3 \pm 1) \times 10^{-5} g_{\mathrm{E}}$, and 64^{+12}_{-20} Pa for the thermal inertia, bulk density, macro-porosity, negative ambient gravity area, peak negative ambient gravity, and cohesive strength, respectively.

[1] Department of Earth and Planetary Sciences, University of Tennessee, Knoxville, Tennessee 37996, USA.

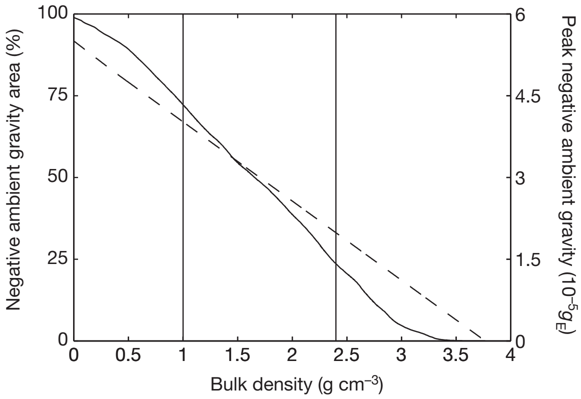


Figure 2 | Degree of negative ambient gravity for (29075) 1950 DA. The area of the surface experiencing negative ambient gravity (solid line) is plotted against the primary (left) y axis, and the peak negative ambient gravity (dashed line) is plotted against the secondary (right) y axis. Both are plotted as functions of bulk density for the nominal diameter of 1.3 km. The vertical lines represent the 1σ range derived for the bulk density, that is, 1.7 ± 0.7 g cm^{-3}.

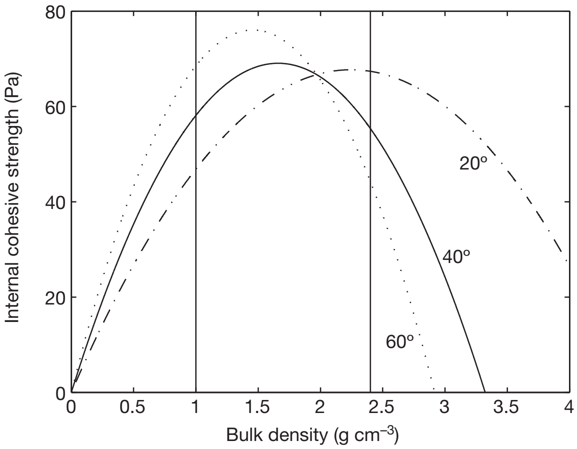


Figure 4 | Minimum internal cohesive strength of (29075) 1950 DA. This was calculated using the Drucker–Prager failure criterion as a function of bulk density (x axis) and angle of friction (shown on the figure) for the nominal diameter of 1.3 km. The vertical lines represent the 1σ range derived for the bulk density, that is, 1.7 ± 0.7 g cm^{-3}.

where rotational centrifugal forces dominate over self-gravity) with peak outward accelerations of $(3 \pm 1) \times 10^{-5} g_E$ (where g_E is 9.81 m s^{-2}) around the equator (Methods; Figs 1, 2 and 3). This makes the presence of a fine-grained regolith unexpected, and requires the existence of cohesive forces for (29075) 1950 DA to retain such a surface. In granular mechanics, the strength of this cohesive force is represented by the bond number B, which is defined as the ratio of this force to the grain's weight. Lunar regolith has been found to be highly cohesive because of van der Waals forces arising between grains[19], and experimental and theoretical studies have shown that the bond number for this cohesive force is given by

$$B = 10^{-5} g_A^{-1} d^{-2} \tag{1}$$

where g_A is the ambient gravity and d is the grain diameter[5]. To prevent loss of surface material requires bond numbers of at least one, but surface stability requires the bond numbers to be greater than ten, which places limits on the possible grain sizes present. For a peak negative ambient gravity of $3 \times 10^{-5} g_E$, this relationship dictates that only grains with diameters less than ~6 cm can be present and stable on the asteroid's surface.

This upper limit of a diameter of ~6 cm for the grains is consistent with (29075) 1950 DA's lunar-like regolith. In particular, lunar regolith has micrometre- to centimetre-sized grains described by an approximate d^{-3} size distribution[6,19]. The rubble-pile asteroid (25143) Itokawa also has a d^{-3} grain size distribution but has boulders ranging up to ~40 m in size on the surface[20], which is reflected in its much higher thermal inertia value of ~750 J m^{-2} K^{-1} s$^{-1/2}$ (ref. 21). (29075) 1950 DA might have had large boulders present on its surface in the past, but these would have been progressively lost in order of size as it was spun-up by the YORP effect (that is, spin state changes caused by the anisotropic reflection and thermal re-emission of sunlight from an irregularly shaped asteroid[10]). This spinning-up selection process leaves behind the relatively fine-grained regolith with low thermal inertia that we infer today[5], and would operate in addition to the thermal fatigue mechanism of asteroid regolith formation[22].

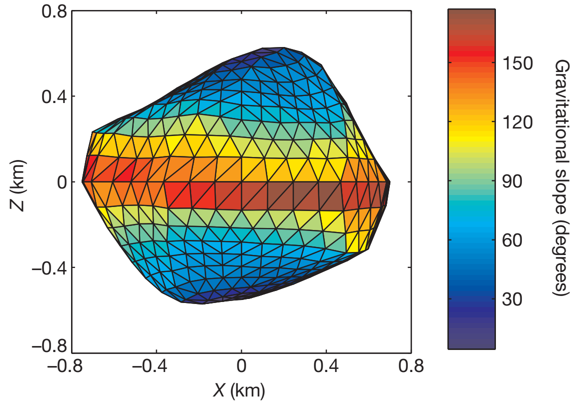


Figure 3 | Gravitational slopes of (29075) 1950 DA. These were produced using the retrograde radar shape model[8] with the nominal derived bulk density of 1.7 g cm^{-3}. Gravitational slopes greater than 90°, which occur predominantly around the equator, indicate that those surface elements are experiencing negative ambient gravity.

To check whether internal cohesive forces are also required to prevent the structural failure of (29075) 1950 DA, we applied the Drucker–Prager model for determining the failure stresses within a geological material[4,6] (Methods). In this model, the maximum spin rate that a rubble-pile asteroid can adopt depends on its overall shape, degree of self-gravity and internal strength. The internal strength results from the angle of friction between constituent grains and any cohesive forces present. Using the dynamically equivalent and equal-volume ellipsoid of (29075) 1950 DA, and using an angle of friction typical for lunar regolith of 40° (ref. 19), we find that a minimum cohesive strength of 64^{+12}_{-20} Pa is required to prevent structural failure (Figs 1 and 4). This is less than that of 100 Pa measured for weak lunar regolith[19], and is within the range of 3–300 Pa estimated by numerical simulations of rubble-pile asteroids[6]. It is also consistent with the range of 40–210 Pa estimated for the precursor body of the main-belt comet P/2013 R3 (ref. 7). This finding proves that not all small asteroids rotating faster than the cohesionless critical spin limit are coherent bodies or monoliths[4–6]. It also supports the view that some high-altitude bursting meteors, such as the impacting asteroid 2008 TC3 (ref. 23), are very small rubble piles held together by cohesive forces[6].

Finally, given that (29075) 1950 DA has a 1 in 19,800 chance of impacting the Earth in 2880 (ref. 12), and has the potential to break up like P/2013 R3 because of its tensional state, there are implications for impact mitigation. Some suggested deflection techniques, such as the kinetic impactor[24], violently interact with the target asteroid and have the potential to destabilize long-ranging granular force networks present[25]. With such tenuous cohesive forces holding one of these asteroids together, a very small impulse may result in complete disruption. This may have happened to the precursor body of P/2013 R3 through a meteorite impact. Therefore, there is a potential danger of turning one Earth-threatening asteroid into several if cohesive forces within rubble-pile asteroids are not properly understood.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 13 May; accepted 26 June 2014.

1. Fujiwara, A. *et al.* The rubble-pile asteroid Itokawa as observed by Hayabusa. *Science* **312,** 1330–1334 (2006).
2. Chesley, S. R. *et al.* Orbit and bulk density of the OSIRIS-REx target asteroid (101955) Bennu. *Icarus* **235,** 5–22 (2014).
3. Walsh, K. J., Richardson, D. C. & Michel, P. Spin-up of rubble-pile asteroids: disruption, satellite formation, and equilibrium shapes. *Icarus* **220,** 514–529 (2012).
4. Holsapple, K. A. Spin limits of Solar System bodies: from the small fast-rotators to 2003 EL61. *Icarus* **187,** 500–509 (2007).
5. Scheeres, D. J., Hartzell, C. M., Sánchez, P. & Swift, M. Scaling forces to asteroid surfaces: the role of cohesion. *Icarus* **210,** 968–984 (2010).
6. Sánchez, P. & Scheeres, D. J. The strength of regolith and rubble pile asteroids. *Meteorit. Planet. Sci.* **49,** 788–811 (2014).
7. Hirabayashi, M., Scheeres, D. J., Sánchez, D. P. & Gabriel, T. Constraints on the physical properties of main belt comet P/2013 R3 from its breakup event. *Astrophys. J.* **789,** L12 (2014).
8. Busch, M. W. *et al.* Physical modeling of near-Earth asteroid (29075) 1950 DA. *Icarus* **190,** 608–621 (2007).
9. Pravec, P. & Harris, A. W. Fast and slow rotation of asteroids. *Icarus* **148,** 12–20 (2000).
10. Bottke, W. F., Vokrouhlický, D., Rubincam, D. P. & Nesvorný, D. The Yarkovsky and YORP effects: implications for asteroid dynamics. *Annu. Rev. Earth Planet. Sci.* **34,** 157–191 (2006).
11. Emery, J. P. *et al.* Thermal infrared observations and thermophysical characterization of OSIRIS-REx target asteroid (101955) Bennu. *Icarus* **234,** 17–35 (2014).
12. Farnocchia, D. & Chesley, S. R. Assessment of the 2880 impact threat from asteroid (29075) 1950 DA. *Icarus* **229,** 321–327 (2014).
13. Rozitis, B. & Green, S. F. Directional characteristics of thermal-infrared beaming from atmosphereless planetary surfaces—a new thermophysical model. *Mon. Not. R. Astron. Soc.* **415,** 2042–2062 (2011).
14. Rozitis, B. & Green, S. F. The influence of rough surface thermal-infrared beaming on the Yarkovsky and YORP effects. *Mon. Not. R. Astron. Soc.* **423,** 367–388 (2012).
15. Mainzer, A. *et al.* NEOWISE observations of near-Earth objects: preliminary results. *Astrophys. J.* **743,** 156 (2011).
16. Rivkin, A. S., Binzel, R. P. & Bus, S. J. Constraining near-Earth object albedos using near-infrared spectroscopy. *Icarus* **175,** 175–180 (2005).
17. Sierks, H. *et al.* Images of asteroid 21 Lutetia: a remnant planetesimal from the early Solar System. *Science* **334,** 487–490 (2011).
18. Wesselink, A. F. Heat conductivity and nature of the lunar surface material. *Bull. Astron. Inst. Netherlands* **10,** 351–360 (1948).
19. Mitchell, J. K., Houston, W. N., Carrier, W. D. & Costes, N. C. Apollo soil mechanics experiment S-200 final report. *Space Sciences Laboratory Series* **15,** 72–85 (Univ. California, Berkeley, 1974); http://www.lpi.usra.edu/lunar/documents/NASA%20CR-134306.pdf.
20. Michikami, T. *et al.* Size-frequency statistics of boulders on global surface of asteroid 25143 Itokawa. *Earth Planets Space* **60,** 13–20 (2008).
21. Müller, T. G., Sekiguchi, T., Kaasalainen, M., Abe, M. & Hasegawa, S. Thermal infrared observations of the Hayabusa spacecraft target asteroid 25143 Itokawa. *Astron. Astrophys.* **443,** 347–355 (2005).
22. Delbo, M. *et al.* Thermal fatigue as the origin of regolith on small asteroids. *Nature* **508,** 233–236 (2014).
23. Jenniskens, P. *et al.* The impact and recovery of asteroid 2008 TC3. *Nature* **458,** 485–488 (2009).
24. Ahrens, T. J. & Harris, A. W. Deflection and fragmentation of near-Earth asteroids. *Nature* **360,** 429–433 (1992).
25. Murdoch, N. *et al.* Simulating regoliths in microgravity. *Mon. Not. R. Astron. Soc.* **433,** 506–514 (2013).

Acknowledgements This publication uses data products from NEOWISE, a project of the Jet Propulsion Laboratory/California Institute of Technology, funded by the Planetary Science Division of NASA. We made use of the NASA/IPAC Infrared Science Archive, which is operated by the Jet Propulsion Laboratory/California Institute of Technology under a contract with NASA. This work was supported by NASA contract NNM10AA11C (Principal Investigator D. S. Lauretta) through the New Frontiers programme, and by NASA contract NNX12AP32G (Principal Investigator J. P. Emery) through the Near-Earth Object Observing programme.

Author Contributions B.R. performed the thermophysical and cohesive force analyses, E.M. retrieved the WISE data and helped with its analysis, and J.P.E. helped with the scientific interpretation of the results. B.R. wrote the manuscript with all co-authors contributing to its final form.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to B.R. (brozitis@utk.edu).

METHODS

Thermophysical modelling. The ATPM was used to determine the thermal inertia and bulk density of (29075) 1950 DA. The ATPM was developed to interpret thermal-infrared observations of planetary surfaces lacking atmospheres[13], and simultaneously make asteroidal Yarkovsky and YORP effect predictions[14]. Accurate interpretation of thermal-infrared observations was verified by applying it to the Moon[13], and it has been successfully applied to asteroids (1862) Apollo[26] and (101955) Bennu[2] to determine their thermal and physical properties.

To summarize how it works[26], the ATPM computes the surface temperature variation for each surface element during a rotation by solving one-dimensional heat conduction with a surface boundary condition that includes direct and multiply scattered sunlight, shadowing, and re-absorbed thermal radiation from interfacing surface elements (that is, global self-heating effects). Rough-surface thermal-infrared beaming (that is, thermal re-emission of absorbed solar energy back towards the Sun) is explicitly included in the form of hemispherical craters, which have been shown to accurately recreate the lunar thermal-infrared beaming effect[13]. The degree of roughness for each surface element is specified by the fraction of its area covered by the (rough) hemispherical craters, f_R. The asteroid thermal emission as a function of wavelength, rotation phase and various thermophysical properties is determined by applying the Planck function to the derived temperatures and summing across visible surface elements. The Yarkovsky and YORP effects are then determined by computing the total recoil forces and torques from photons reflected off and thermally emitted from the asteroid surface.

Analysis of WISE thermal-infrared observations. The thermal inertia of (29075) 1950 DA was determined using archival WISE thermal-infrared observations, which were obtained on 12–13 July 2010 UT (Universal Time) during the WISE All-Sky survey[15]. All instances of WISE observations of (29075) 1950 DA were taken from the Minor Planet Center database and used to query the WISE All-Sky Single Exposure (L1b) source database via the NASA/IPAC Infrared Service Archive. Search constraints of 10′′ within the Minor Planet Center ephemeris of (29075) 1950 DA, and Julian dates within 10 s of the reported observations, were used to ensure proper data retrieval. The magnitudes returned from this query were kept only in the instances in which there was a positive object detection or where a 95% confidence brightness upper limit was reported. (29075) 1950 DA had a faint apparent visual magnitude of 20.5 when the WISE observations were taken, and was only detected at 3σ levels or greater in the W3 (12 μm) and W4 (22 μm) WISE infrared bands. Additionally, we used only data points that repeatedly sampled common rotation phases of (29075) 1950 DA to ensure consistency, and to avoid outliers, within the data set. This resulted in 14 useable data points in the W3 band and 2 useable data points in the W4 band (Extended Data Table 1). The WISE images for these data points were also retrieved to check for any contaminating sources or extended objects surrounding (29075) 1950 DA (see Extended Data Fig. 2). The WISE magnitudes were converted to fluxes, and the reported red–blue calibrator discrepancy[27] was taken into account. A 5% uncertainty was also added in quadrature to the reported observational uncertainties to take into account additional calibration uncertainties[27]. As in previous works of WISE asteroid observations (for example, ref. 15), we colour-correct the model fluxes using the WISE corrections of ref. 27 rather than colour-correcting the observed fluxes.

The free parameters to be constrained by the WISE observations in the model fitting include the diameter, thermal inertia, surface roughness and rotation phase. Although the radar circular polarization ratio indicates a very smooth surface at centimetre-to-decimetre spatial scales[8], it does not provide a constraint on surface roughness occurring at smaller spatial scales that are comparable to the depth of (29075) 1950 DA's thermal skin (~1 mm). Surface roughness occurring at these spatial scales induce the thermal-infrared beaming effect, which requires that roughness must be left as a free parameter to allow the full range of possible interpretations of the WISE thermal-infrared observations to be obtained. In addition, the uncertainty on (29075) 1950 DA's measured rotation period does not allow accurate phasing of the radar shape model between the light-curve observations taken in 2001 and the WISE observations taken in 2010. Therefore, the rotation phase of the first observation (used as the reference) was left as a free parameter in the model fitting.

In the model fitting, the model thermal flux predictions, $F_{\mathrm{MOD}}(\lambda_n, D, \Gamma, f_{\mathrm{R}}, \varphi)$, were compared with the observations, $F_{\mathrm{OBS}}(\lambda_n)$, and observational errors, $\sigma_{\mathrm{OBS}}(\lambda_n)$, by varying the diameter D, thermal inertia Γ, roughness fraction f_{R} and rotation phase φ to give the minimum χ^2 fit

$$\chi^2 = \sum_{n=1}^{N} \left[\frac{F_{\mathrm{MOD}}(\lambda_n, D, \Gamma, f_{\mathrm{R}}, \phi) - F_{\mathrm{OBS}}(\lambda_n)}{\sigma_{\mathrm{OBS}}(\lambda_n)} \right]^2 \quad (2)$$

for a set of $n = 1$ to N observations with wavelength λ_n. Separate thermophysical models were run for thermal inertia values ranging from 0 to 1,000 J m^{-2} K^{-1} s$^{-1/2}$ in equally spaced steps of 20 J m^{-2} K^{-1} s$^{-1/2}$ initially, and then between 0 and 90 J m^{-2} K^{-1} s$^{-1/2}$ in 2 J m^{-2} K^{-1} s$^{-1/2}$ steps once the probable thermal inertia value had been constrained. The diameter, roughness fraction and rotation phase were also stepped through their plausible ranges, forming a four-dimensional grid of model test parameters (or test clones) with the thermal inertia steps. A parameter region bounded by a constant $\Delta\chi^2$ at the 3σ confidence level then defined the range of possible parameters. Finally, counting the number of acceptable test clones in each parameter value bin then allowed us to obtain the probability distribution for each free parameter. The best model fit had a reduced-χ^2 value of 1.06 with a corresponding P value of 0.39, and an example model fit to the WISE observations is shown in Extended Data Fig. 1.

Unfortunately, the WISE data alone do not place unique constraints on the diameter, thermal inertia and surface roughness because of its limited phase angle and wavelength coverage. As shown in Extended Data Fig. 3, the best-fitting diameter increases with thermal inertia such that a unique constraint cannot be made. Fortunately, the radar observations had constrained (29075) 1950 DA's diameter to be 1.3 km with a maximum error of 10% (ref. 8). Therefore, by allowing the diameter to vary between 1.17 km and 1.43 km only, we found that (29075) 1950 DA's thermal inertia value must be very low, that is, 24^{+20}_{-14} J m^{-2} K^{-1} s$^{-1/2}$ or ≤82 J m^{-2} K^{-1} s$^{-1/2}$. This result is consistent with the preliminary upper bound of 110 J m^{-2} K^{-1} s$^{-1/2}$ determined by ref. 28 using a simpler thermophysical model that neglected rough-surface thermal-infrared beaming effects. In our work, the surface roughness still remains unconstrained, but must be included for the Yarkovsky effect analysis described below. The probability distribution for the derived thermal inertia value is shown in Fig. 1.

Analysis of Yarkovsky orbital drift. The bulk density of (29075) 1950 DA could be determined by model-to-measurement comparisons of its Yarkovsky semimajor axis drift. The authors of ref. 12 were able to measure a transverse acceleration of $(-6.70 \pm 1.29) \times 10^{-15}$ AU per day squared acting on (29075) 1950 DA in its orbit by using optical astrometry dating back to 1950 and radar ranging data taken in 2001 and 2012. This transverse acceleration corresponded to a rate of change in semimajor axis of $(-2.95 \pm 0.57) \times 10^{-4}$ AU per million years (Myr) or −44.1 ± 8.5 m yr^{-1} (ref. 29). Parameter studies using the ATPM have shown that the Yarkovsky effect is in general enhanced by rough-surface thermal-infrared beaming[14]. For (29075) 1950 DA's oblate shape, fast rotation period, and low thermal inertia, the potential enhancement was rather large (see Extended Data Fig. 4) and had to be included to prevent underestimation of (29075) 1950 DA's bulk density. The overall Yarkovsky drift acting on (29075) 1950 DA, $\mathrm{d}a/\mathrm{d}t(D,\Gamma,f_{\mathrm{R}},\rho)$, for a bulk density ρ was determined from

$$\frac{\mathrm{d}a}{\mathrm{d}t}(D,\Gamma,f_{\mathrm{R}},\rho) = \left(\frac{D_0}{D}\right)\left(\frac{\rho_0}{\rho}\right)\left[(1-f_{\mathrm{R}})\frac{\mathrm{d}a}{\mathrm{d}t}(\Gamma)_{\mathrm{smooth}} + f_{\mathrm{R}}\frac{\mathrm{d}a}{\mathrm{d}t}(\Gamma)_{\mathrm{rough}} + \frac{\mathrm{d}a}{\mathrm{d}t}(\Gamma)_{\mathrm{seasonal}}\right] \quad (3)$$

where $\mathrm{d}a/\mathrm{d}t(\Gamma)_{\mathrm{smooth}}$ is the smooth surface component, $\mathrm{d}a/\mathrm{d}t(\Gamma)_{\mathrm{rough}}$ is the rough surface component, and $\mathrm{d}a/\mathrm{d}t(\Gamma)_{\mathrm{seasonal}}$ is the seasonal component[26]. Each component was evaluated separately at a specified initial diameter D_0 and bulk density ρ_0. A Yarkovsky effect prediction was produced for every test clone deemed acceptable from the WISE flux-fitting described above. To produce the distribution of possible bulk densities, each prediction was compared against 500 samples of Yarkovsky drift that were randomly selected from a normal distribution with a mean and standard deviation of −44.1 ± 8.5 m yr^{-1}. This ensured that the uncertainty on the measured Yarkovsky drift was taken into account. Extended Data Fig. 3 shows the derived bulk density as a function of thermal inertia for the range of acceptable test clones without a thermal inertia constraint. As indicated, to match the observed drift at the 1σ level required the bulk density to be less than 2.7 g cm^{-3} regardless of the thermal inertia value. Using the thermal inertia constraint, the bulk density was constrained to be 1.7 ± 0.7 g cm^{-3} and its probability distribution is shown in Fig. 1.

Cohesive force modelling. Surface and internal cohesive forces are required to prevent surface mass shedding and structural failure of (29075) 1950 DA, respectively. The surface cohesive forces are proportional to the magnitude of the negative ambient gravity experienced by the surface. In particular, the gravitational acceleration $\boldsymbol{g}$ acting at a particular point of (29075) 1950 DA's surface $\boldsymbol{x}$ was determined using a polyhedral gravity field model[30]. Under asteroid rotation the ambient gravitational acceleration at that surface point will be modified by centripetal acceleration, such that

$$\boldsymbol{g}' = \boldsymbol{g} - \omega^2\left(\boldsymbol{x} - \left(\boldsymbol{x}\cdot\hat{\boldsymbol{p}}\right)\hat{\boldsymbol{p}}\right) \quad (4)$$

where ω is the asteroid angular velocity and $\hat{\boldsymbol{p}}$ is the unit vector specifying the orientation of the asteroid rotation pole. The ambient surface gravity g_{A} acting along the surface normal $\hat{\boldsymbol{n}}$ of point $\boldsymbol{x}$ is then given by

$$g_{\mathrm{A}} = -\boldsymbol{g}'\cdot\hat{\boldsymbol{n}} \quad (5)$$

and, finally, the effective gravitational slope θ is given by

$$\theta = \cos^{-1}(g_A/|\mathbf{g}'|) \tag{6}$$

To accurately measure the area of the asteroid's surface experiencing negative ambient gravity each shape model facet was split into one hundred smaller subfacets. The negative ambient gravity as a function of bulk density is shown in Fig. 2, and a three-dimensional plot of gravitational slope is shown in Fig. 3. The bond number for a particular regolith grain diameter in a specified degree of negative ambient gravity was then determined from equation (1).

The minimum internal cohesive force required to prevent structural failure of (29075) 1950 DA was determined analytically from the Drucker–Prager failure criterion and a model of interior stresses within the asteroid[4,6]. For a homogenous ellipsoidal body with semi-axes a, b and c the average normal stress components are

$$\overline{\sigma}_x = \left(\rho\omega^2 - 2\pi\rho^2 G A_x\right)\frac{a^2}{5}$$

$$\overline{\sigma}_y = \left(\rho\omega^2 - 2\pi\rho^2 G A_y\right)\frac{b^2}{5}$$

$$\overline{\sigma}_z = \left(-2\pi\rho^2 G A_z\right)\frac{c^2}{5} \tag{7}$$

where G is the gravitational constant. The A_i terms are dimensionless coefficients that depend only on the shape of the body, and are $A_x = 0.57003$, $A_y = 0.60584$, and $A_z = 0.82413$ for the dynamically equivalent and equal-volume ellipsoid of (29075) 1950 DA, which were determined from equation (4.6) of ref. 4. The Drucker–Prager failure criterion using the average stresses is given by

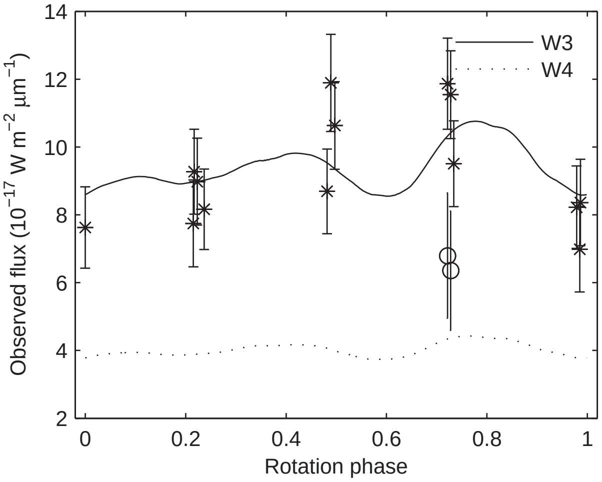
$$\frac{1}{6}\left[\left(\overline{\sigma}_x - \overline{\sigma}_y\right)^2 + \left(\overline{\sigma}_y - \overline{\sigma}_z\right)^2 + \left(\overline{\sigma}_z - \overline{\sigma}_x\right)^2\right] \le \left[k - s\left(\overline{\sigma}_x + \overline{\sigma}_y + \overline{\sigma}_z\right)\right]^2 \tag{8}$$

where k is the internal cohesion and s is the slope constant. The slope constant is determined from the angle of friction φ using

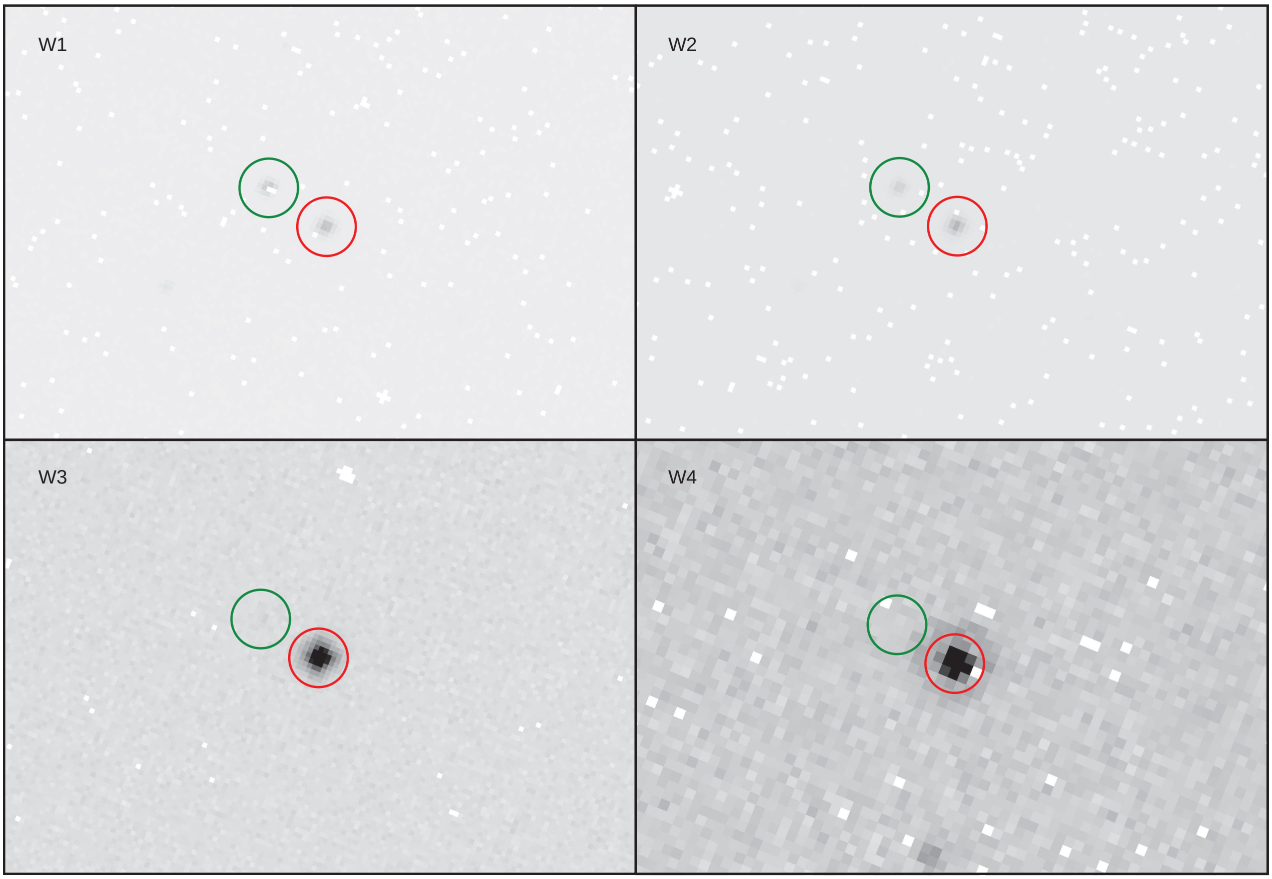
$$s = \frac{2\sin\varphi}{\sqrt{3}(3 - \sin\varphi)} \tag{9}$$

An angle of friction consistent with lunar regolith of 40° (ref. 19) was assumed to calculate the minimum internal cohesive force required to prevent structural failure of (29075) 1950 DA using the Drucker–Prager criterion. The minimum internal cohesive force as a function of bulk density for three different angles of friction is shown in Fig. 4.

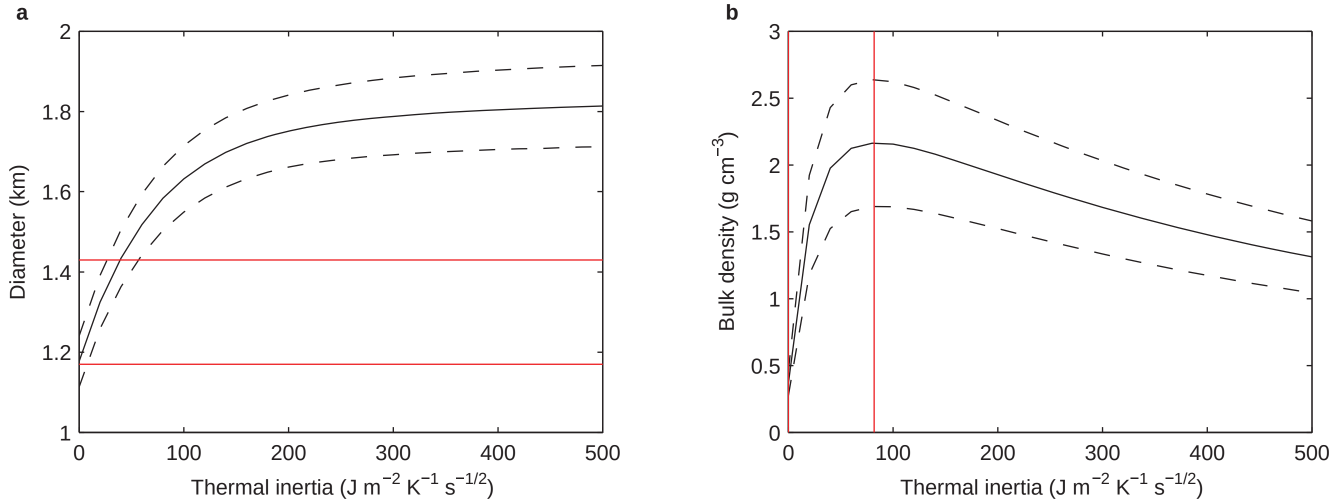
26. Rozitis, B., Duddy, S. R., Green, S. F. & Lowry, S. C. A thermophysical analysis of the (1862) Apollo Yarkovsky and YORP effects. *Astron. Astrophys.* **555,** A20 (2013).
27. Wright, E. L. *et al.* The Wide-field Infrared Survey Explorer (WISE): mission description and initial on-orbit performance. *Astron. J.* **140,** 1868–1881 (2010).
28. Nugent, C. R. *Solar Radiation and near-Earth asteroids: Thermophysical Modeling and New Measurements of the Yarkovsky Effect.* 3556842, http://search.proquest.com/docview/1328407433, PhD thesis, Univ. California, Los Angeles (ProQuest, UMI Dissertations Publishing, 2013).
29. Farnocchia, D. *et al.* Near Earth asteroids with measurable Yarkovsky effect. *Icarus* **224,** 1–13 (2013).
30. Werner, R. A. & Scheeres, D. J. Exterior gravitation of a polyhedron derived and compared with harmonic and mascon gravitation representations of asteroid 4769 Castalia. *Celestial Mech. Dyn. Astron.* **65,** 313–344 (1997).



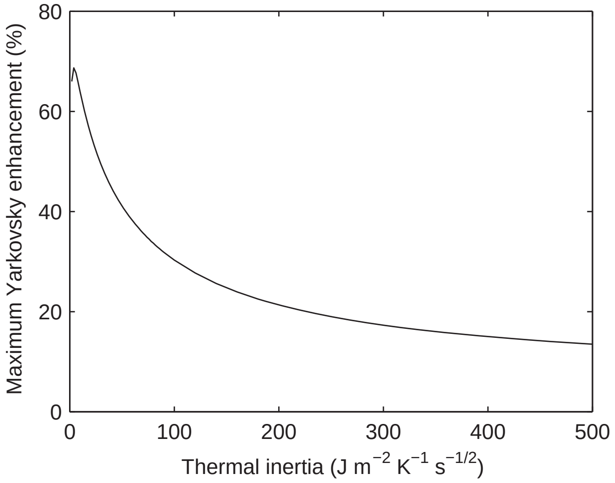
Extended Data Figure 1 | Example ATPM fit to the WISE thermal-infrared observations. This fit (lines for WISE infrared bands W3 and W4) was made for a thermal inertia of 24 J m^{-2} K^{-1} s$^{-1/2}$ and a surface roughness of 50%. The error bars correspond to the 1σ uncertainties on the measured data points.



Extended Data Figure 2 | WISE thermal-infrared images of (29075) 1950 DA. The image scale is 2.75 arcsec per pixel for the W1, W2 and W3 bands and 5.53 arcsec per pixel for the W4 band. White pixels are 'bad' pixels that do not contain data. The object seen to the upper left of (29075) 1950 DA (red circle) in the W1 (3.4 μm) and W2 (4.6 μm) bands is a faint background star (green circle).



Extended Data Figure 3 | Physical properties derived for (29075) 1950 DA as a function of thermal inertia. **a**, Diameter; **b**, bulk density. The dashed lines represent the 1σ uncertainty for the average solid lines. The red horizontal lines represent the radar diameter constraint[8] of 1.30 ± 0.13 km, and the red vertical lines represent the corresponding thermal inertia constraint of $\leq$82 J m^{-2} K^{-1} s$^{-1/2}$.



Extended Data Figure 4 | Enhancement of Yarkovsky orbital drift by surface roughness for (29075) 1950 DA.

Extended Data Table 1 | WISE thermal-infrared observations of (29075) 1950 DA

MJD* (day)	Rotation phase	WISE infrared band	WISE magnitude	Flux (10^{-17} W m^{-2} μm^{-1})	Heliocentric distance (AU)	WISE-centric distance (AU)	Phase angle (°)
55389.71982	0.000	W3	10.00 ± 0.16	7.63 ± 1.20	1.738	1.410	35.8
55389.85212	0.497	W3	9.64 ± 0.12	10.64 ± 1.29	1.739	1.409	35.8
55390.11670	0.490	W3	9.52 ± 0.12	11.89 ± 1.43	1.741	1.408	35.7
55390.18282	0.238	W3	9.93 ± 0.15	8.17 ± 1.19	1.741	1.407	35.7
55390.24890	0.985	W3	10.10 ± 0.19	6.98 ± 1.26	1.741	1.407	35.7
55390.24903	0.987	W3	9.90 ± 0.16	8.36 ± 1.27	1.741	1.407	35.7
55390.31510	0.734	W3	9.76 ± 0.13	9.51 ± 1.27	1.742	1.407	35.7
55390.38121	0.482	W3	9.86 ± 0.15	8.69 ± 1.25	1.742	1.406	35.7
55390.51351	0.978	W3	9.92 ± 0.15	8.23 ± 1.22	1.743	1.406	35.7
55390.57973	0.727	W3	9.55 ± 0.11	11.54 ± 1.30	1.744	1.405	35.7
55390.71200	0.224	W3	9.82 ± 0.15	8.98 ± 1.28	1.745	1.405	35.6
55390.84433	0.721	W3	9.52 ± 0.11	11.87 ± 1.34	1.745	1.404	35.6
55390.97650	0.216	W3	9.98 ± 0.17	7.75 ± 1.28	1.746	1.403	35.6
55390.97660	0.217	W3	9.79 ± 0.14	9.27 ± 1.25	1.746	1.403	35.6
55390.57973	0.727	W4	7.16 ± 0.30	6.35 ± 1.76	1.744	1.405	35.7
55390.84433	0.721	W4	7.09 ± 0.29	6.80 ± 1.85	1.745	1.404	35.6

* MJD is Modified Julian Day, that is, MJD = JD − 2400000.5.

Extended Data Table 2 | Physical properties of (29075) 1950 DA

	Property	Value
Size	Diameter of equivalent volume sphere[8]	1.30 ± 0.13 km
	Dimensions of dynamically-equivalent and equal-volume ellipsoid (2*a*, 2*b*, 2*c*)[8]	1.46 × 1.39 × 1.07 km
Optical	Absolute magnitude[8]	16.8 ± 0.2
	Phase parameter[8]	0.15 ± 0.10
	Geometric albedo[8]	0.20 ± 0.05
Rotation	Rotation period[8]	2.12160 ± 0.00004 hr
	Obliquity[8]	168 ± 5 °
Orbit	Semimajor axis[12]	1.70 AU
	Eccentricity[12]	0.51
	Yarkovsky semimajor axis drift[12]	(-2.95 ± 0.57) × 10^{-4} AU/Myr (or -44.1 ± 8.5 m yr^{-1})
Surface composition	Spectral type[16]	M
	Thermal inertia*	24 $^{+20}/_{-14}$ J m^{-2} K^{-1} s$^{-1/2}$ (36 $^{+30}/_{-20}$ J m^{-2} K^{-1} s$^{-1/2}$ at 1 AU)
	Surface roughness*	50 ± 30 %
	Radar albedo[8]	0.23 ± 0.05
	Radar circular polarization ratio[8]	0.14 ± 0.03
Mass	Bulk density*	1.7 ± 0.7 g cm^{-3}
	Macro-porosity*	51 ± 19 %
	Mass*	(2.1 ± 1.1) × 10^{12} kg
Cohesion	Surface area of negative ambient gravity*	48 ± 24 %
	Peak negative ambient gravity*	(3 ± 1) × 10^{-5} g_E
	Internal cohesive strength*	64 $^{+12}/_{-20}$ Pa

* Derived in this work.

doi:10.1038/nature13617

Formation of monatomic metallic glasses through ultrafast liquid quenching

Li Zhong[1], Jiangwei Wang[1], Hongwei Sheng[2,3], Ze Zhang[4] & Scott X. Mao[1]

It has long been conjectured that any metallic liquid can be vitrified into a glassy state provided that the cooling rate is sufficiently high[1–4]. Experimentally, however, vitrification of single-element metallic liquids is notoriously difficult[5]. True laboratory demonstration of the formation of monatomic metallic glass has been lacking. Here we report an experimental approach to the vitrification of monatomic metallic liquids by achieving an unprecedentedly high liquid-quenching rate of $10^{14}\,\mathrm{K\,s^{-1}}$. Under such a high cooling rate, melts of pure refractory body-centred cubic (bcc) metals, such as liquid tantalum and vanadium, are successfully vitrified to form metallic glasses suitable for property interrogations. Combining *in situ* transmission electron microscopy observation and atoms-to-continuum modelling, we investigated the formation condition and thermal stability of the monatomic metallic glasses as obtained. The availability of monatomic metallic glasses, being the simplest glass formers, offers unique possibilities for studying the structure and property relationships of glasses. Our technique also shows great control over the reversible vitrification–crystallization processes, suggesting its potential in micro-electromechanical applications. The ultrahigh cooling rate, approaching the highest liquid-quenching rate attainable in the experiment, makes it possible to explore the fast kinetics and structural behaviour of supercooled metallic liquids within the nanosecond to picosecond regimes.

Since the first discovery of metallic glass (MG) in the 1960s (ref. 1), the search for new types of MG has not stopped[2,6,7]. To date, most MG formers are known to consist of two or more elements with distinct atomic sizes and chemical affinities[2,8], usually formed by quenching the liquids with techniques varying from conventional die casting[6] (10^{1}–$10^{3}\,\mathrm{K\,s^{-1}}$), melt spinning[2] ($10^{5}$–$10^{6}\,\mathrm{K\,s^{-1}}$), liquid splat-quenching[9] ($\sim10^{9}$–$10^{10}\,\mathrm{K\,s^{-1}}$) to pulsed laser quenching[10] ($\sim10^{12}$–$10^{13}\,\mathrm{K\,s^{-1}}$). New techniques such as nanocalorimetry[11] (10^{4}–$10^{6}\,\mathrm{K\,s^{-1}}$) have also emerged, vying for high heating and cooling rates. Unfortunately, these solidification techniques can hardly be applied to the production of monatomic MGs, mainly because of the extremely low glass-forming ability of monatomic metallic liquids, resulting from their fast nucleation and crystal growth kinetics at deep undercoolings[4,12–14]. Thus, vitrification of pure monatomic MGs requires extremely high critical cooling rates, far above the experimentally accessible level, to suppress crystal growth. The monatomic MG may also be confronted with the thermal stability issue at room temperature, at which spontaneous crystallization seems inevitable[14].

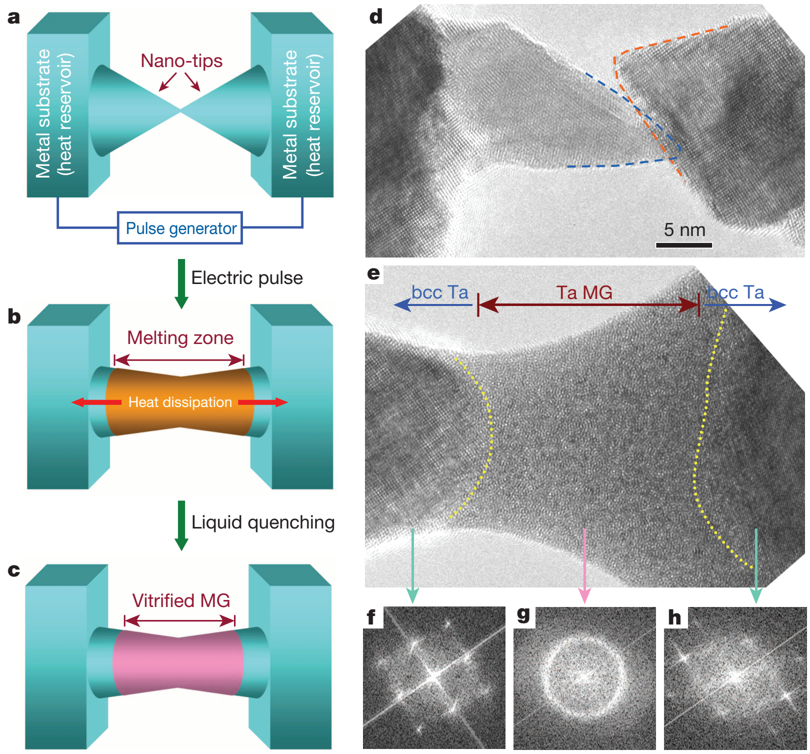


Figure 1 | Illustration of an ultrafast liquid-quenching approach. **a**–**c**, Schematic drawing of the experimental configuration. Two protruded nano-tips are brought into contact with each other (**a**) and are melted by the application of a short square electric pulse with a duration of ~3.7 ns and a voltage in the range 0.5–3 V (**b**). Heat dissipates rapidly through the two bulk substrates (indicated by two red arrows), vitrifying the melting zone to form monatomic MGs (**c**). **d**, **e**, High-resolution TEM images showing two contacting Ta nano-tips (**d**) forming a Ta MG (**e**) after the application of a 0.8-V, 3.6-ns electric pulse. The GCIs are denoted by yellow dotted curves. **f**–**h**, Fast Fourier transformations confirming a fully vitrified region 20 nm long and 15 nm thick (**g**) bounded by two crystalline substrates viewed along the ⟨100⟩ (**f**) and ⟨110⟩ (**h**) crystallographic orientations, respectively.

[1] Department of Mechanical Engineering and Materials Science, University of Pittsburgh, Pittsburgh, Pennsylvania 15261, USA. [2] School of Physics, Astronomy and Computational Sciences, George Mason University, Fairfax, Virginia 22030, USA. [3] Center for High Pressure Science and Technology Advanced Research, Shanghai 201203, China. [4] Department of Materials Science and Engineering and State Key Laboratory of Silicon Materials, Zhejiang University, Hangzhou 310027, China.

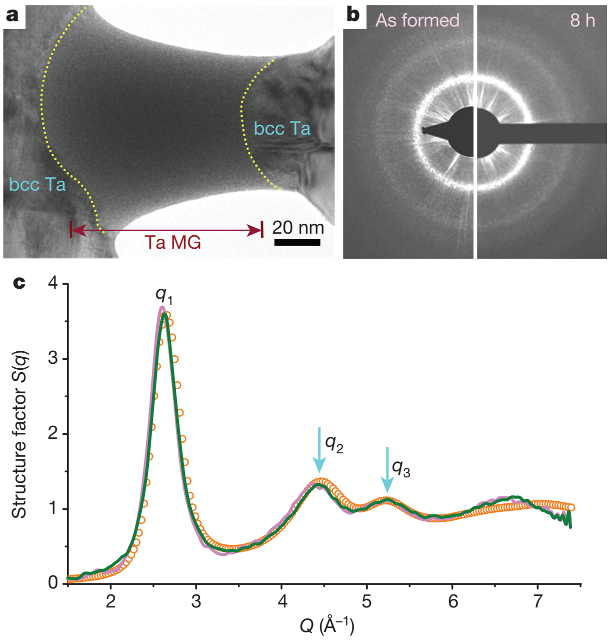


Figure 2 | Structure and thermal stability of Ta MG. **a**, TEM morphology of a typical Ta MG with length of ~90 nm and diameter of ~60 nm. The GCIs are indicated by yellow dotted curves. **b**, Electron diffraction of Ta MG, as quenched (left) and after relaxation for ~8 h (right). **c**, Comparison of the structure factors of the Ta MGs as formed (pink line), after relaxation for 8 h (green line) and simulated (orange circles). All three curves show very similar peak positions, including well-separated second (q_2) and third (q_3) peaks (indicated by cyan arrows). The ratios of peak positions are the same for the relaxed and simulated structures: $q_2/q_1 = 1.68$ and $q_3/q_1 = 1.99$.

Consequently, except for a few special circumstances (for example at very thin edges of a splat-quenched nickel foil[9]), monatomic MGs have not been found to form from pure metal melts by vitrification.

More recently, pure metallic germanium liquid was reported to vitrify under hydrostatic pressure above 7.9 GPa (ref. 5). However, on releasing pressure to ambient condition, germanium MG quickly transforms to a non-metallic low-density amorphous phase, in which case the tendency and mechanism of liquid vitrification are largely different from those of most *d*-block transition metals. Other non-vitrification methods (such as vapour deposition[15] and chemical synthesis[16]) have been attempted to produce monatomic amorphous samples, which are often in geometrically confined forms (such as substrate-supported thin films and nano-sized powders) and are plagued with either purity[16] or stability[17] problems, offering limited potential for broader applications. Therefore advanced techniques to fabricate high-purity monatomic MGs with controllable geometries are highly appealing. By building an *in situ* Joule heating nanodevice inside a transmission electron microscope (TEM), we present a unique ultrafast liquid-quenching system for vitrifying monatomic metallic liquids. This technique exploits the excellent thermal conductivity of the metals and maximizes the heat conduction rate of the cooling system.

Our ultrafast quenching technique is illustrated in Fig. 1a–c (see Methods). First, two protruded nano-tips with clean surfaces (Extended Data Fig. 1a, b) are brought into contact with each other (Fig. 1a) under an ultrahigh vacuum condition inside the TEM. A short square electric pulse, typically 0.5–3 V in amplitude and within 3.7 ns in duration, imposes local Joule heating on the joined tips, causing melting of the extrusion tips and the formation of a melting zone in the middle (Fig. 1b). On instantaneous cessation of the electric pulse and, consequently, local Joule heating, heat dissipates rapidly through the solidifying piece and the conductive heat reservoir, creating an extremely high cooling rate sufficient to vitrify the melt into a glassy state (Fig. 1c). In Fig. 1d, e we demonstrate that a 0.8-V, 3.6-ns electric pulse on two connecting crystalline Ta nano-tips (Fig. 1d) led to the formation of a Ta MG 15 nm wide and 20 nm long (Fig. 1e). Structural characterization of the MG is presented below. The dimensions of the MGs can be controlled by tuning electric pulse parameters while engaging *in situ* tensile or compressive loading. In this way, Ta MGs with dimensions of 100 nm in diameter or an aspect ratio of ~4 are obtainable (Extended Data Fig. 1c, d). The formation of even larger Ta MGs, which are not electron transparent, was not pursued in this work. Applying this method, we have systematically tested the vitrification capability of transition metals and successfully obtained Ta, V, W and Mo monatomic MGs; their morphologies are provided in Extended Data Figs 2–4. The materials systems that have been vitrified are typically early transition bcc metals with high melting points and excellent thermal conductivities.

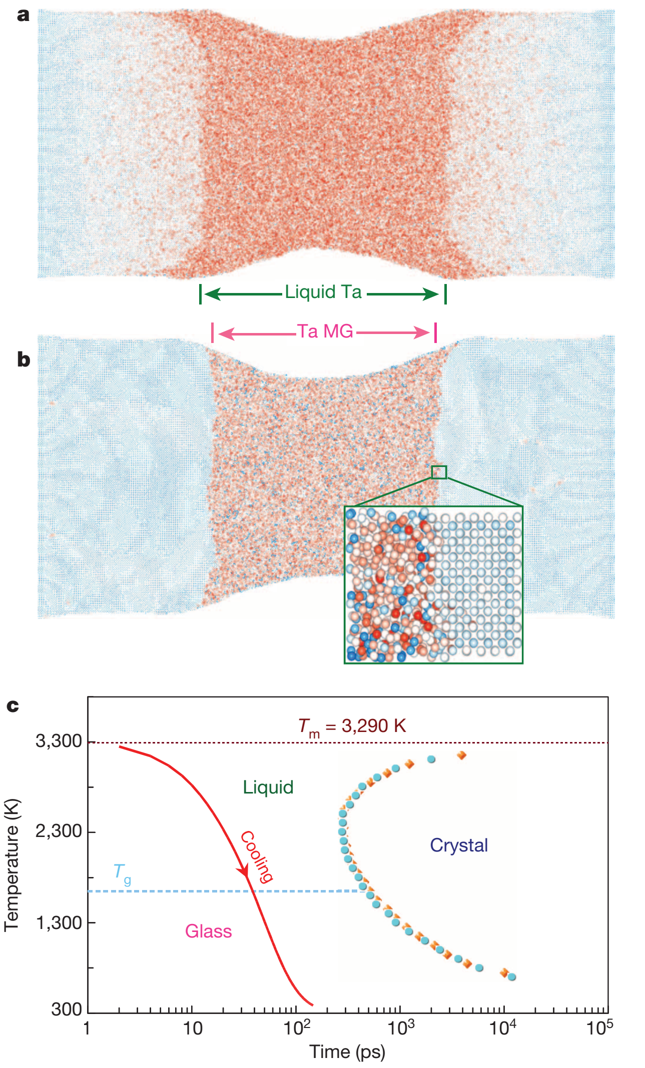


Figure 3 | Dynamic vitrification process in liquid Ta revealed by AtC computer simulation. **a**, Atomic configuration showing a liquid zone of Ta 35 nm in length after Joule heating (at $t = 0$ ps). The atoms are coloured on the basis of their degree of disorder represented by local bond-orientational order parameter q_6 (ref. 29). The red colour corresponds to liquid Ta after Joule heating. **b**, Atomic configuration showing the formation of a Ta MG segment 30 nm in length after quenching ($t = 150$ ps). The average temperature of the Ta nanowire is close to room temperature at $t = 150$ ps. The inset highlights the interface structure between amorphous and bcc Ta. **c**, A time–temperature–transformation diagram derived from isothermal MD simulations, outlining approximately the formation condition of Ta MG. The crystal zone is estimated from the crystal growth rates of the (100) plane (cyan circles) and the (110) plane (orange squares). The red solid line indicates the temperature evolution of the moving LCI (and later on GCI) during cooling.

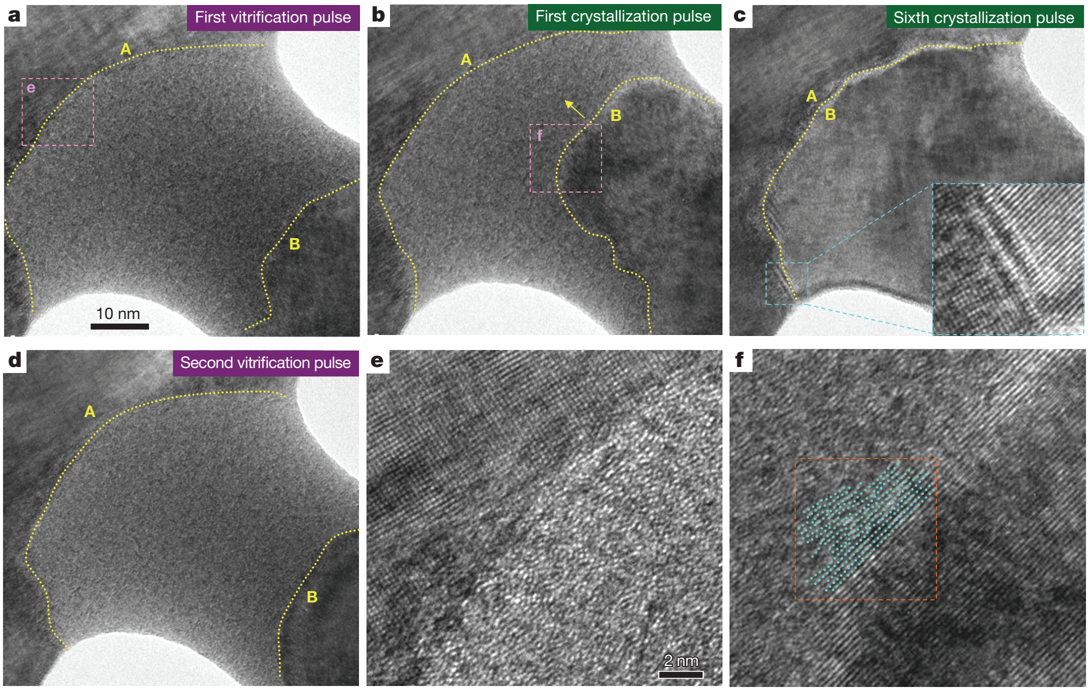

a
First vitrification pulse
A
e
B
10 nm
b
First crystallization pulse
A
B
f
c
Sixth crystallization pulse
A
B
d
Second vitrification pulse
A
B
e
2 nm
f

Figure 4 | Reversible crystallization–vitrification phase changes of Ta MG. **a**, Formation of a Ta MG 40 nm thick and 50 nm long under a 3.6-ns, 1.26-V electric pulse. The two GCIs are indicated by yellow dotted curves and are labelled A and B, respectively. **b**, **c**, Controlled gradual crystallization under a series of pulses 3.6 ns in duration and 0.90 V in amplitude. Crystallization proceeded with crystal growth at GCI B (indicated by a yellow arrow) and completed after six crystallization pulses (inset in **c**). **d**, A second vitrification pulse resulted in the formation of a Ta MG similar to that shown in **a**. **e**, **f**, Close-up views of the atomically rough and diffuse GCIs during a phase-change cycle. A schematic drawing with cyan dotted lines along one set of the (110) planes shows the gradual breakdown of the long-range order across the GCI.

The amorphous nature of the MG as obtained was confirmed by TEM diffraction patterns. As a first check, the diffusive diffraction halos in the fast Fourier transformation (Fig. 1g) of the area bounded by two glass–crystal interfaces (GCIs) are characteristic of amorphous structure, contrasting the bright diffraction spots of the Ta substrates with a well-defined bcc structure (Fig. 1f, h). To confirm the glassy structure and the thermal stability of Ta MG, a sample 60 nm in diameter and 90 nm in length was relaxed in high vacuum at room temperature for 8 h (Fig. 2a). Electron diffraction patterns of the Ta MGs as quenched (Fig. 2b, left) and after relaxation (Fig. 2b, right) showed similar features characterized by diffuse halos typical of amorphous structure. The corresponding integrated and optimized one-dimensional static structure factors[18] $S(q)$ (Fig. 2c) showed similarities in their shape and peak positions, indicating that no major structural changes occurred in Ta MG after 8 h. The slight shift to the right in the main peak positions of $S(q)$ may be attributed to structural relaxation in the glass, as expected. The main peak positions of the relaxed Ta MG were measured to be 2.63 Å^{-1}, 4.42 Å^{-1} and 5.23 Å^{-1}, corresponding to $q_2/q_1 = 1.68$ and $q_3/q_1 = 1.99$, which are almost identical to the simulated structure factor (orange circles in Fig. 2c) derived by quenching liquid Ta at a cooling rate of $\sim 10^{13}$ K s^{-1} on the computer. The observed $S(q)$ of Ta MG also agrees well with theoretical works on monatomic systems[19–22], as well as with previous experimental results on amorphous cobalt[23] and iron[24], for which $q_2/q_1 = 1.69$ and $q_3/q_1 = 1.97$.

To understand the vitrification process of the liquid and estimate the cooling rate, atoms-to-continuum (AtC) simulations[25] have been performed, where the molecular dynamics (MD) system is coupled to an additional electron temperature field that implements the two-temperature model (TTM)[25] for heat transport (see Methods). The present experimental approach is capable of vitrifying monatomic metallic liquids on temporal and spatial scales commensurate with those in MD modelling, permitting a direct comparison between experiment and MD simulation and enabling accurate interpretation of the multiphysics of the cooling process. Quenching of liquid Ta starts at the moment when external Joule heating is turned off (Extended Data Fig. 5a), during which the temperature evolution in the nanowire depends on rapid heat dissipation through the massive crystalline substrates kept at room temperature. As a result of the large temperature gradient, excellent heat conductivity and small specimen size, ultrafast cooling is achieved, as demonstrated by the evolution of the temperature distribution in the Ta nanowire (Extended Data Fig. 5b). The computed cooling rate of the liquid zone (Extended Data Fig. 5c) reaches as high as 10^{14} K s^{-1} at 4,200 K and decreases slightly to 5×10^{13} K s^{-1} at the glass transition temperature T_g of liquid Ta, which is estimated to be $\sim$1,650 K (Extended Data Fig. 6).

Accompanying the rapid quenching process, the real-time dynamics of the atomic system was revealed by MD simulations, indicating that whether a MG can eventually form is determined by the competition between the liquid-quenching rate and the crystal growth rate from the melt. Under the given experimental condition, a large portion of the original liquid Ta zone 35 nm in length was vitrified after Joule heating was cut off (Fig. 3a, b), demarcated by atomically rough GCIs (Fig. 3b inset), corroborating our experimental observations. More generally, the time needed for complete crystallization due to crystal growth at the (100) (cyan circles) and the (110) (orange squares) interfaces of the pre-existing crystals (that is, the crystalline substrates) at different temperatures is computed and plotted in the time–temperature–transformation diagram for Ta (Fig. 3c), based on the growth rates of these two liquid–crystal interfaces (LCIs) from the melt (Extended Data Fig. 5d). In the AtC simulation shown in Fig. 3a, b, the temperature evolution of the LCI (and later on the GCI) of the system follows the red curve in Fig. 3c, which trends into the glass-forming region (that is, the left side of the time–temperature–transformation curves) and corresponds to a quenching rate one order of magnitude higher than the critical cooling rate of $\sim 5 \times 10^{12}$ K s^{-1} estimated from the dimensional consideration (described in Methods), justifying the formation of Ta MG under the present experimental configuration. The effect of a trailing edge in the applied electric pulses is taken into account by modelling liquid quenching under a heat flux terminated within 0.4 ns in a ramp function rather than instantly.

As shown in Extended Data Fig. 7a, b, 18 nm of the 35-nm Ta liquid was successfully vitrified into MG under a cooling rate varying between 3×10^{13} and 10^{13} K s^{-1} (Extended Data Fig. 7c).

The thermal stability of Ta MG is rationalized by our computation, showing that the crystal growth of low-index faces of bcc Ta is a thermally activated process at room temperature, with infinitesimally small growth rates at the GCIs (based on Extended Data Fig. 5d). It should be pointed out that the 'slow' growth rate of Ta crystals is distinctly different from that of face-centred cubic (fcc) metals, in which the growth of crystal interfaces is expected to be spontaneous and fast even at zero temperature[14]. Indeed, we have tried but failed to produce any monatomic MGs from fcc metals (for example gold, silver, copper, palladium, aluminium, rhodium and iridium) using the very same approach.

The competition between vitrification and crystal growth can be controlled experimentally, leading to a novel phase-change phenomenon in the MG. Figure 4 illustrates a vitrification–crystallization cycle in a Ta sample controlled by alternately applying two kinds of electric pulse with the same duration (3.6 ns) but different voltages. With the assistance of *in situ* TEM observation, the structural and morphological evolutions of the sample can be monitored on the fly. As shown in Fig. 4a, a Ta MG 40 nm in thickness and 50 nm in length obtained with a high-voltage (1.26-V) electric pulse (that is, the vitrification pulse) reverted to its original crystalline state after the application of a series of low-voltage (0.90-V) electric pulses (that is, the crystallization pulses), with each pulse reducing the size of the sandwiched amorphous zone (Fig. 4b, c). The GCIs were identified as being atomically rough and diffuse during both vitrification (Fig. 4e) and crystallization (Fig. 4f). After complete crystallization of the Ta MG (Fig. 4c and inset), another vitrification pulse again generated a glassy zone of Ta (Fig. 4d) with almost identical dimensions and morphology to that shown in Fig. 4a, demonstrating that a reversible glass–crystal phase-change process is achievable by our approach. Another example of controlled phase changes in Ta MG is presented in Supplementary Video 1. This reversible phase-change behaviour of Ta, bearing a resemblance to those in chalcogenide glasses[26,27], indicates the potential of the present methodology for employing marginal glass formers with extremely fast crystallization kinetics for applications in phase-change-based nano-devices[26,27].

The vitrification of pure metallic liquids reported here should not be attributed to the enhanced glass-forming ability associated with impurities in the original materials (Extended Data Table 1) and/or oxygen contamination during experimental procedures (see Methods and Extended Data Fig. 8). The successful formation of monatomic MGs opens up new opportunities for studying the structural dependence of rheological, thermal, electrical and mechanical properties of MGs, in which complications due to chemical effects in multicomponent MGs can be shielded. For instance, we have conducted tensile tests on the monatomic Ta MG as synthesized (Extended Data Fig. 9 and Supplementary Video 2). The insight gained from the mechanical testing experiment is beyond the scope of the present paper and will be presented later. Last, we stress that the ultrafast quenching rate ($\sim 10^{14}$ K s^{-1}) achieved in this technique is high enough to freeze the atoms within a fraction of a nanosecond. With such a high cooling rate to reach deep quench, the inherent structure of liquids[28] can be accessed, enabling investigations to be made of the fast kinetics of supercooled liquids and the mechanisms for the formation of metastable materials under conditions far from equilibrium.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 6 April; accepted 25 June 2014.
Published online 6 August 2014.

1. Klement, W., Willens, R. H. & Duwez, P. Non-crystalline structure in solidified gold–silicon alloys. *Nature* **187,** 869–870 (1960).
2. Greer, L. A. Metallic glasses. *Science* **267,** 1947–1953 (1995).
3. Cohen, M. H. & Turnbull, D. Molecular transport in liquids and glasses. *J. Chem. Phys.* **31,** 1164–1169 (1959).
4. Turnbull, D. Under what conditions can a glass be formed? *Contemp. Phys.* **10,** 473–488 (1969).
5. Bhat, M. H. *et al.* Vitrification of a monatomic metallic liquid. *Nature* **448,** 787–790 (2007).
6. Johnson, W. L. Fundamental aspects of bulk metallic glass formation in multicomponent alloys. *Mater. Sci. Forum* **225–227,** 35–50 (1996).
7. Ding, S. Y. *et al.* Combinatorial development of bulk metallic glasses. *Nature Mater.* **13,** 494–500 (2014).
8. Johnson, W. L. Thermodynamic and kinetic aspects of the crystal to glass transformation in metallic materials. *Prog. Mater. Sci.* **30,** 81–134 (1986).
9. Davies, H. A., Aucote, J. & Hull, J. B. Amorphous nickel produced by splat quenching. *Nature* **246,** 13–14 (1973).
10. Lin, C.-J. & Spaepen, F. Fe-B glasses formed by picosecond pulsed laser quenching. *Appl. Phys. Lett.* **41,** 721–723 (1982).
11. McCluskey, P. J. & Vlassak, J. J. Combinatorial nanocalorimetry. *J. Mater. Res.* **25,** 2086–2100 (2010).
12. Aga, R. S., Morris, J. R., Hoyt, J. J. & Mendelev, M. Quantitative parameter-free prediction of simulated crystal-nucleation times. *Phys. Rev. Lett.* **96,** 245701 (2006).
13. Trudu, F., Donadio, D. & Parrinello, M. Freezing of a Lennard–Jones fluid: from nucleation to spinodal regime. *Phys. Rev. Lett.* **97,** 105701 (2006).
14. Broughton, J. Q., Gilmer, G. H. & Jackson, K. A. Crystallization rates of a Lennard–Jones liquid. *Phys. Rev. Lett.* **49,** 1496–1500 (1982).
15. Fujime, S. Electron diffraction at low temperature. II. Radial distribution analysis of metastable structure of metal films prepared by low temperature condensation. *Jpn. J. Appl. Phys.* **5,** 778–787 (1966).
16. Suslick, K. S., Choe, S.-B., Cichowlas, A. A. & Grinstaff, M. W. Sonochemical synthesis of amorphous iron. *Nature* **353,** 414–416 (1991).
17. Hilsch, R. in *Non-Crystalline Solids* (ed. Fréchette, V. D.) 348–373 (Wiley, 1960).
18. Sheng, H. W. *et al.* Polyamorphism in a metallic glass. *Nature Mater.* **6,** 192–197 (2007).
19. Ichikawa, T. The assembly of hard spheres as a structure model of amorphous iron. *Phys. Status Solidi A* **29,** 293–302 (1975).
20. Sachdev, S. & Nelson, D. R. Theory of the structure factor of metallic glasses. *Phys. Rev. Lett.* **53,** 1947–1950 (1984).
21. Yamamoto, R., Matsuoka, H. & Doyama, M. Structural relaxation of the dense random packing model for amorphous iron. *Phys. Status Solidi A* **45,** 305–314 (1978).
22. Dzugutov, M. Glass formation in a simple monatomic liquid with icosahedral inherent local order. *Phys. Rev. A* **46,** R2984–R2987 (1992).
23. Leung, P. K. & Wright, J. G. Structural investigations of amorphous transition element films. I. Scanning electron diffraction study of cobalt. *Phil. Mag.* **30,** 185–194 (1974).
24. Ichikawa, T. Electron diffraction study of the local atomic arrangement in amorphous iron and nickel films. *Phys. Status Solidi A* **19,** 707–716 (1973).
25. Jones, R. E., Templeton, J. A., Wagner, G. J., Olmsted, D. & Modine, N. A. Electron transport enhanced molecular dynamics for metals and semi-metals. *Int. J. Numer. Methods Eng.* **83,** 940–967 (2010).
26. Siegrist, T. *et al.* Disorder-induced localization in crystalline phase-change materials. *Nature Mater.* **10,** 202–207 (2011).
27. Wuttig, M. & Yamada, N. Phase-change materials for rewriteable data storage. *Nature Mater.* **6,** 824–832 (2007).
28. Stillinger, F. H. & Weber, T. A. Packing structures and transitions in liquids and solids. *Science* **225,** 983–989 (1984).
29. Steinhardt, P. J., Nelson, D. R. & Ronchetti, M. Bond-orientational order in liquids and glasses. *Phys. Rev. B* **28,** 784–805 (1983).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank A. J. Coley and Y. Liu for their assistance in designing the electrical circuit for our experiment, and Y. He for performing the liquid-quenching experiment on iridium and rhodium. H.W.S. acknowledges helpful discussions with L. V. Zhigilei. S.X.M. acknowledges National Science Foundation (NSF) grant CMMI 0928517 through the support of the University of Pittsburgh. Work at George Mason University was supported by the US NSF under grant no. DMR-0907325. This work was performed in part at the Center for Integrated Nanotechnologies (CINT), an Office of Science User Facility operated for the US Department of Energy (DOE) Office of Science. The computational work was conducted on the SR10000-K1/52 supercomputing facilities of the Institute for Materials Research, Tohoku University.

Author Contributions L.Z. and J.W.W. carried out the TEM experiments under the direction of S.X.M. S.X.M. created the experimental protocols. H.W.S. performed the computer simulation. L.Z., H.W.S. and S.X.M. performed the experimental data analysis. All the authors (L.Z., J.W.W., H.W.S., Z.Z. and S.X.M.) contributed to the discussion of the results. L.Z., S.X.M. and H.W.S. wrote the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to S.X.M. (sxm2@pitt.edu) or H.W.S. (hsheng@gmu.edu).

doi:10.1038/nature13639

The tidal–rotational shape of the Moon and evidence for polar wander

Ian Garrick-Bethell[1,2], Viranga Perera[1]†, Francis Nimmo[1] & Maria T. Zuber[3]

The origin of the Moon's large-scale topography is important for understanding lunar geology[1], lunar orbital evolution[2] and the Moon's orientation in the sky[3]. Previous hypotheses for its origin have included late accretion events[4], large impacts[5], tidal effects[6] and convection processes[7]. However, testing these hypotheses and quantifying the Moon's topography is complicated by the large basins that have formed since the crust crystallized. Here we estimate the large-scale lunar topography and gravity spherical harmonics outside these basins and show that the bulk of the spherical harmonic degree-2 topography is consistent with a crust-building process controlled by early tidal heating throughout the Moon. The remainder of the degree-2 topography is consistent with a frozen tidal–rotational bulge that formed later, at a semi-major axis of about 32 Earth radii. The probability of the degree-2 shape having both tidal-heating and frozen shape characteristics by chance is less than 1%. We also infer that internal density contrasts eventually reoriented the Moon's polar axis by 36 ± 4°, to the configuration we observe today. Together, these results link the geology of the near and far sides, and resolve long-standing questions about the Moon's large-scale shape, gravity and history of polar wander.

The theory of equilibrium figures of rotating fluid bodies is a classic problem in geophysics, and it has been helpful in understanding the shapes of the Sun and planets. However, the origin of the Moon's shape has remained an open problem in the past century[2,6,8–10], and the body's deviations from any simple tidal–rotational (spherical harmonic degree-2) figure are large[11]. This difficulty is surprising given the Moon's presumably simple early thermal history: born hot and quickly cooled, one might expect the Moon to be described by a simple figure of equilibrium.

Researchers have traditionally suggested that the Moon's degree-2 spherical harmonic gravity coefficients, which have been used as proxies for the degree-2 shape, are especially large when compared to higher-degree coefficients[9,12]. Figure 1 shows a power law or 'Kaula's rule' fit to degrees $n = 3$ to 50 for the Moon's gravity[13] and topography data[14]. The power at degree 2 is 4.5 times and 2.6 times the power expected from extrapolating the best-fit power law, for gravity and topography, respectively, supporting the idea that the degree-2 coefficients are unique. Indeed, the fraction of excess power for topography is greater than the excesses for Venus, Earth and Mars (Supplementary Information). The Moon's strong degree-2 power has been interpreted as a frozen tidal–rotational state inherited from when the Moon was closer to the Earth; this is known as the fossil bulge hypothesis[6]. An open problem, however, has been that the ratio of the $C_{2,0}$ and $C_{2,2}$ spherical harmonic coefficients is different from the expected value by a factor of 2.6 (refs 2 and 10).

Adding to the fossil bulge idea and motivated by tidal processes in Europa's ice shell[15], Garrick-Bethell *et al.*[16] inferred that the farside highlands crust has a degree-2 shape that is explainable by tidal heating during the magma ocean epoch. However, ref. 16 did not address the rest of the Moon's shape, the Moon's orientation history, and the details of gravity and topography when they are examined together. In particular, ref. 16 did not reconcile its results with the classic fossil bulge hypothesis[10], or explain its anomalous $C_{2,0}/C_{2,2}$ ratio. To address these problems and create a unified explanation for the Moon's degree-2 shape and orientation, we here consider two effects: the Moon's largest basins, and the reference frame in which we analyse lunar topography.

The South Pole–Aitken basin (SPA) is the largest[17], deepest[1] and oldest lunar basin[5], and its degree-2 power affects our interpretation of the primordial degree-2 shape. In addition to SPA, we focus on the 12 largest basins that produce obvious local anomalies in topography, crustal thickness or gravity (in all, 22% of the surface, Fig. 2a–c). To determine the Moon's degree-2 shape without these basins, we fit spherical harmonics of degrees $n = 0$ to 5 to data outside their boundaries. Figure 2d and f shows the Moon's topography and appearance after rotation to the reference frame where the only non-zero degree-2 terms are $C_{2,0}$ and $C_{2,2}$ (with $C_{2,0} < 0$), hereafter termed 'the principal frame'. If the Moon's outer figure, as opposed to its internal density distribution, once controlled the lunar moments of inertia (see below), this would be the reference frame that once faced Earth. This frame's largest principal axis is at (6 ± 4° S, 30 ± 1° E), its polar axis is at (54 ± 5° N, 309 ± 6° E), and its intermediate axis is at (35.1 ± 5° S, 296.4 ± 4° E) (Fig. 2a–c).

Without the largest basins, the Moon's topography power spectrum displays substantially less variance at low degrees. Performing a power-law fit for $n = 3$ to 50 using the new power at degrees 3, 4 and 5 (Fig. 1,

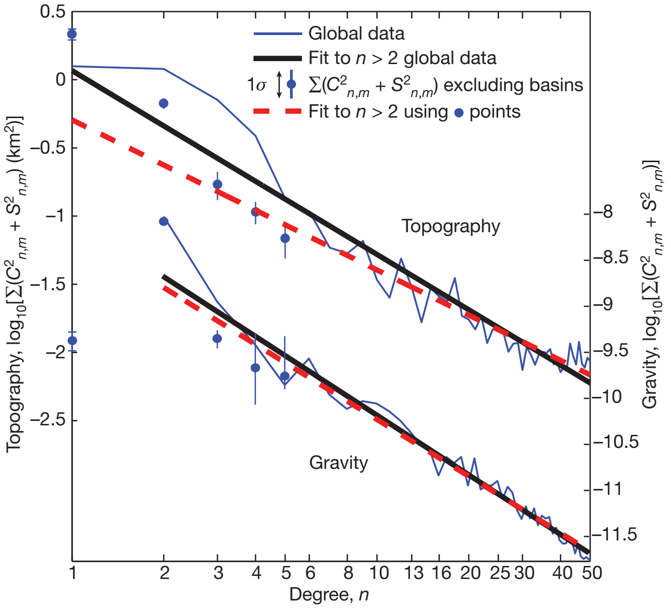


Figure 1 | Lunar topography and gravity power spectra, with best-fit power laws for degrees $n = 3$ to 50. The blue dots show the power using data outside large basins (±1σ). The blue dot at degree-1 for gravity is due to a small displacement of the lunar centre of mass when the basins are removed. See Supplementary Information section 1.

[1]Department of Earth and Planetary Sciences, University of California, Santa Cruz, 1156 High Street, Santa Cruz, California 95064, USA. [2]School of Space Research, Kyung Hee University, 1 Seocheon-dong, Giheung-gu, Yongin-si, Gyeonggi-do 446-701, Korea. [3]Department of Earth, Atmospheric and Planetary Sciences, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, Massachusetts 02139, USA. †Present address: School of Earth and Space Exploration, Arizona State University, PO Box 876004, Tempe, Arizona 85287-6004, USA.

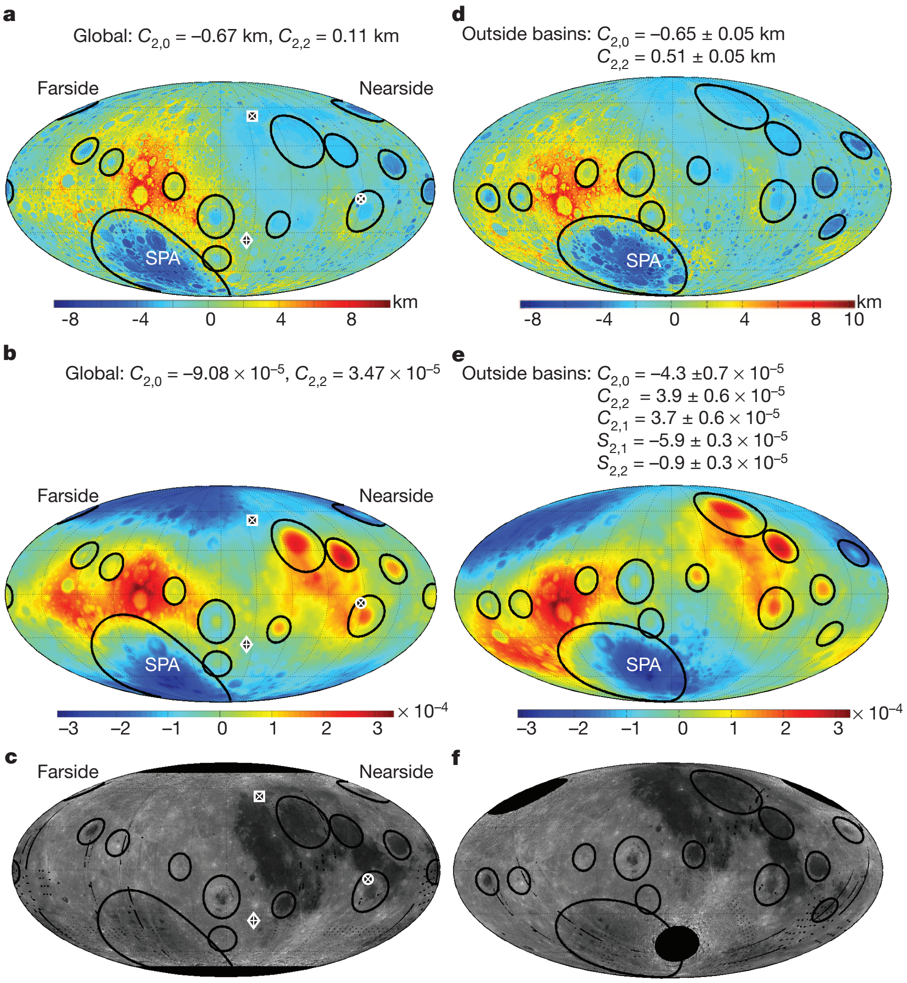

a
Global: $C_{2,0}$ = –0.67 km, $C_{2,2}$ = 0.11 km
Farside
Nearside
SPA
km
-8
-4
0
4
8
d
Outside basins: $C_{2,0}$ = –0.65 ± 0.05 km
$C_{2,2}$ = 0.51 ± 0.05 km
SPA
km
-8
-4
0
4
8
10
b
Global: $C_{2,0} = -9.08 \times 10^{-5}$, $C_{2,2} = 3.47 \times 10^{-5}$
Farside
Nearside
SPA
× 10–4
–3
–2
–1
0
1
2
3
e
Outside basins: $C_{2,0} = -4.3 \pm 0.7 \times 10^{-5}$
$C_{2,2} = 3.9 \pm 0.6 \times 10^{-5}$
$C_{2,1} = 3.7 \pm 0.6 \times 10^{-5}$
$S_{2,1} = -5.9 \pm 0.3 \times 10^{-5}$
$S_{2,2} = -0.9 \pm 0.3 \times 10^{-5}$
SPA
× 10–4
–3
–2
–1
0
1
2
3
c
Farside
Nearside
f

Figure 2 | The topography, gravity and appearance of the Moon, with black lines illustrating basins removed in the analysis. a, Lunar topography. The crossed black circle, diamond and square in **a**, **b** and **c** are the primordial minimum, intermediate and maximum principal moment of inertia axes, respectively. **b**, Expansion of degree-2 to degree-360 lunar gravity potential coefficients (multiply by $2.823 \times 10^6 \mathrm{m}^2 \mathrm{s}^{-2}$ to obtain the surface potential). **c**, Lunar 750-nm spectral reflectance, with the data above 75° latitude blacked out. **d**, The data in **a** after rotation to the topography principal frame, using rotation angles calculated from data outside large basins. **e**, The data in **b** after rotating to the topography principal frame, as in **d**. **f**, The data in **c** after rotation to the principal topography frame, as in **d**.

dashed red line), we find the degree-3 and degree-4 power is much closer to the predictions from Kaula's rule. However, the degree-2 power remains in excess by a factor of 2.8. The Moon's strong degree-2 power, even without its large basins, implies that purely local explanations for the degree-2 character of the far side, such as a late-accreting second moon[4], are less plausible.

To address the origin of the Moon's primordial degree-2 shape, we must also consider the degee-2 gravity potential of the Moon (Fig. 2b). If we again fit degree-2 coefficients outside the basins, we find that gravity's largest principal axis shifts only 5 ± 2°, from (0° N, 180° E) to (5 ± 2° S, 182 ± 1° E), and its polar axis only 5 ± 2° to (85 ± 2° N, 203 ± 35° E) (Supplementary Table 7). In addition, the degree-2 gravity power decreases by a small amount, 12% (Fig. 1, blue dot). The weak effect of basins on the degree-2 gravity potential is partly due to SPA's nearly compensated state[18], and SPA's large contribution (45%) to the area removed.

The gravity and topography principal frame calculations above reveal a previously unappreciated but critical problem in understanding the lunar shape: while both gravity and topography have anomalously high degree-2 power, the principal topography and gravity reference frames do not align at present (that is, using global data), and nor do they align when using degree-2 harmonics fitted outside the largest basins. Using global data, the largest gravity and topography principal axes are separated by 34°, and using data outside large basins, the largest principal axes are separated by 30° ± 5°. Therefore, other non-basin events distorted the Moon from any single, simple equilibrium figure in either gravity or topography, making it unclear which data set represents the primordial frame where any tidal-rotational effects were acquired.

However, a simple argument suggests that topography's principal frame formed first. Degree-2, tidally produced crustal thickness variations[16], if they exist, must have developed early when the lithosphere was weak enough to permit significant tidal flexing, and will therefore be isostatically compensated (with a relatively small gravity signature). Furthermore, any uncompensated fossil component of shape, if it exists, must have frozen-in after the lithosphere cooled and strengthened, and degree-2 crustal thickness growth largely ceased. Therefore, as long as the crustal thickness variations produced degree-2 topography that dominated any subsequent fossil topography, and the principal axes remained mostly fixed while forming, topography's principal frame will be the Moon's first-established Earth-oriented principal frame. Below, we will demonstrate that topography components from both crustal-thickening (compensated) and fossil-bulge (uncompensated) processes probably exist in topography's principal frame, with the crustal component being larger, and that each topography component has the $C_{2,0}/C_{2,2}$ ratio expected from each unique process.

To assess the nature of the degree-2 topography in the primordial, basin-removed principal topography frame, we examine the associated

Table 1 | Compensated and uncompensated degree-2 topography harmonics

	Compensated topography (±1σ)	Uncompensated topography (±1σ)
$C_{2,0}$	−0.53 ± 0.07 km	−0.11 ± 0.04 km
$C_{2,2}$	0.40 ± 0.06 km	0.11 ± 0.03 km
$C_{2,0}/C_{2,2}$	−1.3 ± 0.2	−1.0 ± 0.3

Solution for the combination of compensated and uncompensated topography to match the $C_{2,0}$ and $C_{2,2}$ gravity and topography harmonics shown in Fig. 2d and e. Compensated and uncompensated topography are associated with crustal thickness variations and a frozen fossil bulge, respectively. The solution assumes a crustal density 2,550 kg m^{-3}, mantle density 3,200 kg m^{-3}, mean lunar density 3,340 kg m^{-3}, and a mean crustal thickness of 40 km (ref. 20). $C_{2,0}$ values do not sum exactly to −0.65 km because of rounding.

gravity harmonics in the same frame (Fig. 2e). In this frame, we use a joint analysis of gravity and topography to find that neither completely compensated nor completely uncompensated topography alone can explain the $C_{2,0}$ and $C_{2,2}$ gravity coefficients (Supplementary Information section 4). However, in Table 1 we show that a linear combination of compensated and uncompensated topography is consistent with gravity and topography observations; the topography is effectively about 80% compensated (shown graphically in Supplementary Fig. 9).

Having established that both fossil (uncompensated) and crustal thickness (compensated) topography components are required, we can examine their coefficient ratios to test their origins. The ratio of $C_{2,0}/C_{2,2}$ for normalized gravity and topography coefficients is −0.96 (which is approximately −1.0) for frozen tidal–rotational fossil bulges in low-eccentricity synchronous orbits (and assuming that the normalized polar moment of inertia is 0.4)[10,12,19]. The classic problem has been that the observed present-frame ratio is very different from −1.0: it is −2.6 for global gravity[2,10] (Fig. 2b) and −6.1 for global topography (Fig. 2a). However, we must now also consider the expected topography ratio for tidally controlled crustal thickness variations[16]. Unlike the case for fossil topography, this ratio is variable, depending on the amount of tidal dissipation. Although dissipation depends on a number of parameters that are difficult to estimate, such as lower crustal viscosity, we find that for 114 model calculations spanning a variety of conditions, $C_{2,0}/C_{2,2}$ approaches −1.1 to −1.3 as the mean global tidal heat flux increases above about 50 mW m^{-2} (Fig. 3).

From Table 1, we see the ratio $C_{2,0}/C_{2,2}$ for compensated topography in topography's principal frame is −1.3 ± 0.2, and for uncompensated topography, the ratio is −1.0 ± 0.3. These values are consistent with the ratios predicted for a crust sculpted by tidal heating, and a frozen fossil bulge, respectively. A similar spherical harmonic coefficient fit to a model of crustal thickness[20], with large basins removed, yields $C_{2,0}/C_{2,2}$ = −1.1 ± 0.2 (Supplementary Fig. 10), in good agreement with the compensated topography ratio. The observed topography $C_{2,0}/C_{2,2}$ ratios are robust (compared to their uncertainties) to the inclusion or exclusion of different basins, as well as increases in the size of SPA up to 50% (30% for other basins), and changes in the maximum fit degree (Supplementary Table 4). If we had not removed the effects of large basins, the solution for compensated and uncompensated topography in the global-topography principal frame yields $C_{2,0}/C_{2,2}$ values of −2.0 and −6.1, respectively (Supplementary Table 10).

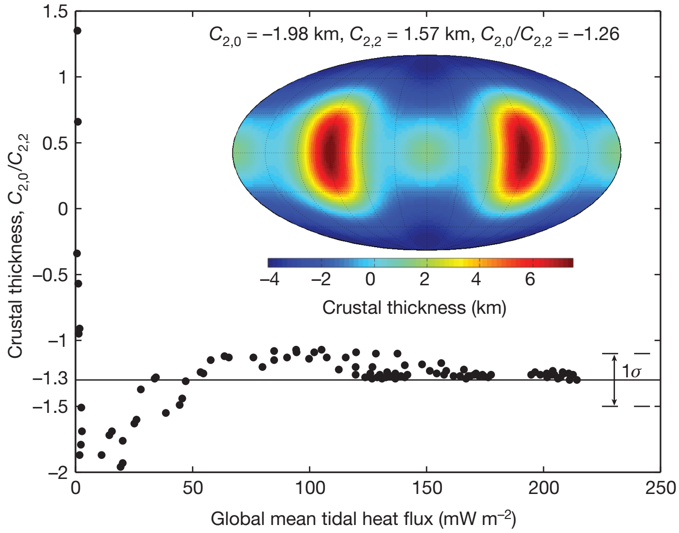


Figure 3 | The ratio $C_{2,0}/C_{2,2}$ for crustal thickness (or compensated topography), as a function of global mean tidal heat flux, for 114 model cases. (See Supplementary Table 13.) The observed ratio of −1.3 ± 0.2 (1σ, dashed lines) for compensated topography outside of large basins is illustrated (Table 1). The inset shows a model crustal thickness map with $C_{2,0}/C_{2,2}$ = −1.26 (Supplementary Information section 8).

To assess the likelihood that the unique $C_{2,0}/C_{2,2}$ ratios arise by chance, we performed Monte Carlo simulations with topography and gravity with the same statistical properties as the observed data. This topography and gravity could arise from any source, including early mantle convection processes, or the process that produced the Moon's centre-of-mass/centre-of-figure offset. We find that the probabilities of the compensated and uncompensated topography $C_{2,0}/C_{2,2}$ ratios randomly falling between −1.1 to −1.3, and between −0.9 to −1.1 (ranges taken to represent the predicted values for each mechanism), are 8% and 5%, respectively (Supplementary Information section 9 and Supplementary Figs 12 and 13). The joint probability is only 0.3%, suggesting that the degree-2 shape is tidally produced.

In the principal topography frame, we also obtain gravity terms $S_{2,1}$, $C_{2,1}$ and $S_{2,2}$, which constitute 59% of the basin-removed degree-2 gravity power (Fig. 2e). Since these terms are associated with zero topography, they arise from subsurface density anomalies that must have developed after a rigid lithosphere formed. Dynamically produced hemisphere-scale density changes have been proposed[21–23], and these would probably have degree-2 power that could have affected the Moon's degree-2 tidal signatures. We can estimate the probability that the Moon's tidal characteristics would survive such changes. For example, starting with just the $C_{2,0}$ and $C_{2,2}$ gravity and topography values for the basin-removed Moon, a randomly placed hemisphere-sized gravity anomaly that yields the same total degree-2 gravity power as the basin-removed Moon permits survival (<30% alteration) of the compensated topography $C_{2,0}/C_{2,2}$ ratio 92% of the time, and survival of the uncompensated ratio 37% of the time (Supplementary Information section 10). This simple model demonstrates that if the Moon's unique tidal signatures form (which is seldom by chance; see above), their recovery is quite plausible despite subsequent internal gravity changes. This is largely because the $C_{2,0}/C_{2,2}$ ratios are dependent on topography, not gravity alone.

Our tidal calculations indicate that when the semi-major axis a exceeds about 25 Earth radii (R_E), no realistic models can produce significant tidal heating, and when a is less than about 10R_E, the orbital evolution timescales (less than a million years) are too short to have built a significant amount of crust. The uncompensated $C_{2,0}$ and $C_{2,2}$ values imply fossil freeze-in at $a \approx 32R_E$ or $a \approx 30R_E$ allowing for 18% relaxation after four billion years[24]. This freeze-in location is larger than 25R_E (above), and therefore consistent with the requirement that the lithosphere must have formed after the crust-building epoch. The location is also consistent with freeze-in before the Cassini state transition ($a \approx$ 30–34R_E)[25], which would have affected the lunar shape. Nominally, it takes roughly 200–300 million years for the Moon to evolve to $a \approx 32R_E$ after accretion[26]. This lithosphere development timescale is consistent with estimates of 100–200 million years for complete magma ocean crystallization, based on radioisotope studies and thermal modelling[27–29]. By combining timescales such as these, and our inferred fossil formation position at $a \approx 32R_E$, the orbital evolution and tidal properties of the early Earth–Moon system can be further constrained.

Finally, we find the principal topography frame places the Moon's palaeopole in northern Oceanus Procellarum (54 ± 5° N, 309 ± 6° E), and about 30° from the centre of the thorium-rich Procellarum KREEP (enriched in potassium, calcium and the rare-earth elements) terrane (Supplementary Fig. 14). This palaeopole location may be testable by using the poles of magnetized portions of the crust[30]. Eventually, the additional gravity in $C_{2,1}$, $S_{2,1}$ and $S_{2,2}$, plus the basins we have removed, changed the lunar moments of inertia, and reoriented the Moon to the present frame we see today. While the details and timing of these later processes are not yet fully understood, a self-consistent origin of the primordial degree-2 shape helps to provide a framework for understanding the many subsequent events in lunar evolution.

Received 27 February; accepted 1 July 2014.

Published online 30 July 2014.

1. Zuber, M. T., Smith, D. E., Lemoine, F. G. & Neumann, G. A. The shape and internal structure of the Moon from the Clementine Mission. *Science* **266,** 1839–1843 (1994).
2. Jeffreys, H. On the figures of the Earth and Moon. *Geophys. J. Int.* **4,** 1–13 (1937).
3. Melosh, H. J. Large impact craters and the moon's orientation. *Earth Planet. Sci. Lett.* **26,** 353–360 (1975).
4. Jutzi, M. & Asphaug, E. Forming the lunar farside highlands by accretion of a companion moon. *Nature* **476,** 69–72 (2011).
5. Wilhelms, D. E. The geologic history of the Moon. *USGS Prof. Paper 1348* (US Government Printing Office, 1987).
6. Sedgwick, W. F. On the figure of the Moon. *Messenger Math.* **27,** 171–173 (1898).
7. Loper, D. E. & Werner, C. L. On lunar asymmetries 1. Tilted convection and crustal asymmetry. *J. Geophys. Res.* **107,** http://dx.doi.org/10.1029/2000JE001441 (2002).
8. Urey, H. C., Elsasser, W. M. & Rochester, M. G. Note on the internal structure of the Moon. *Astrophys. J.* **129,** 842–848 (1959).
9. Lambeck, K. & Pullan, S. The lunar fossil bulge hypothesis revisited. *Phys. Earth Planet. Inter.* **22,** 29–35 (1980).
10. Stevenson, D. J. Origin and implications of the degree two lunar gravity field. *Proc. Lunar Sci. Conf.* **32,** 1175 (2001).
11. Smith, D. E., Zuber, M. T., Neumann, G. A. & Lemoine, F. G. Topography of the Moon from the Clementine LIDAR. *J. Geophys. Res.* **102,** 1591–1611 (1997).
12. Williams, J. G., Boggs, D. H., Yoder, C. F., Ratcliff, J. T. & Dickey, J. O. Lunar rotational dissipation in solid body and molten core. *J. Geophys. Res.* **106,** 27933–27968 (2001).
13. Zuber, M. T. *et al.* Gravity field of the Moon from the gravity recovery and interior laboratory (GRAIL) mission. *Science* **339,** 668–671 (2013).
14. Smith, D. E. *et al.* Initial observations from the Lunar Orbiter Laser Altimeter (LOLA). *Geophys. Res. Lett.* **37,** L18204 (2010).
15. Ojakangas, G. W. & Stevenson, D. J. Thermal state of an ice shell on Europa. *Icarus* **81,** 220–241 (1989).
16. Garrick-Bethell, I., Nimmo, F. & Wieczorek, M. A. Structure and formation of the lunar farside highlands. *Science* **330,** 949–951 (2010).
17. Garrick-Bethell, I. & Zuber, M. T. Elliptical structure of the lunar South Pole-Aitken basin. *Icarus* **204,** 399–408 (2009).
18. Namiki, N. *et al.* Farside gravity field of the Moon from four-way Doppler measurements of SELENE (Kaguya). *Science* **323,** 900–905 (2009).
19. Garrick-Bethell, I., Wisdom, J. & Zuber, M. T. Evidence for a past high eccentricity lunar orbit. *Science* **313,** 652–655 (2006).
20. Wieczorek, M. A. *et al.* The crust of the Moon as seen by GRAIL. *Science* **339,** 671–675 (2013).
21. Zhong, S., Parmentier, E. M. & Zuber, M. T. A dynamic origin for the global asymmetry of lunar mare basalts. *Earth Planet. Sci. Lett.* **177,** 131–140 (2000).
22. Laneuville, M., Wieczorek, M. A., Breuer, D. & Tosi, N. Asymmetric thermal evolution of the Moon. *J. Geophys. Res.* **118,** 1435–1452 (2013).
23. Melosh, H. J. Mascons and the moon's orientation. *Earth Planet. Sci. Lett.* **25,** 322–326 (1975).
24. Zhong, S. & Zuber, M. T. Long-wavelength topographic relaxation for self-gravitating planets and implications for the time-dependent compensation of surface topography. *J. Geophys. Res.* **105,** 4153–4164 (2000).
25. Siegler, M. A., Bills, B. G. & Paige, D. A. Effects of orbital evolution on lunar ice stability. *J. Geophys. Res.* **116,** E03010 (2011).
26. Dwyer, C. A., Stevenson, D. J. & Nimmo, F. A long-lived lunar dynamo driven by continuous mechanical stirring. *Nature* **479,** 212–214 (2011).
27. Solomon, S. C. & Longhi, J. Magma oceanography: 1. Thermal evolution. *Proc. Lunar Sci. Conf.* **8,** 583–599 (1977).
28. Borg, L. E., Connelly, J. N., Boyet, M. & Carlson, R. W. Chronological evidence that the Moon is either young or did not have a global magma ocean. *Nature* **477,** 70–72 (2011).
29. Meyer, J., Elkins, L. T. & Wisdom, J. Coupled thermal–orbital evolution of the early Moon. *Icarus* **208,** 1–10 (2010).
30. Runcorn, S. K. Lunar magnetism, polar displacements and primeval satellites in the Earth–Moon system. *Nature* **304,** 589–596 (1983).

Supplementary Information is available in the online version of the paper.

Acknowledgements This work was supported by the BK21-plus programme through the National Research Foundation (NRF), funded by the Ministry of Education of Korea. Conversations with E. Mazarico are appreciated. We acknowledge the GRAIL and LRO teams for the gravity and topography data used in the analysis.

Author Contributions I.G.-B. and M.T.Z. planned the research. V.P. performed the power-law calculations and helped develop the spherical harmonic fitting procedures. F.N. performed the tidal heating calculations. I.G.-B. performed the remainder of the research and wrote the paper, with contributions from all authors.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to I.G.-B. (igarrick@ucsc.edu).

LETTER

doi:10.1038/nature13402

Neuropsychosocial profiles of current and future adolescent alcohol misusers

Robert Whelan[1,2], Richard Watts[3], Catherine A. Orr[4], Robert R. Althoff[5,6], Eric Artiges[7,8], Tobias Banaschewski[9], Gareth J. Barker[10], Arun L. W. Bokde[11], Christian Büchel[12,13], Fabiana M. Carvalho[10], Patricia J. Conrod[10,14], Herta Flor[9], Mira Fauth-Bühler[9,15], Vincent Frouin[16], Juergen Gallinat[12,17], Gabriela Gan[18], Penny Gowland[19], Andreas Heinz[17], Bernd Ittermann[20], Claire Lawrence[21], Karl Mann[9], Jean-Luc Martinot[7,22], Frauke Nees[9], Nick Ortiz[1,23], Marie-Laure Paillère-Martinot[17,22], Tomas Paus[24,25], Zdenka Pausova[26], Marcella Rietschel[9], Trevor W. Robbins[27], Michael N. Smolka[18], Andreas Ströhle[17], Gunter Schumann[10,28], Hugh Garavan[1,6,11] & the IMAGEN Consortium†

A comprehensive account of the causes of alcohol misuse must accommodate individual differences in biology, psychology and environment, and must disentangle cause and effect. Animal models[1] can demonstrate the effects of neurotoxic substances; however, they provide limited insight into the psycho-social and higher cognitive factors involved in the initiation of substance use and progression to misuse. One can search for pre-existing risk factors by testing for endophenotypic biomarkers[2] in non-using relatives; however, these relatives may have personality or neural resilience factors that protect them from developing dependence[3]. A longitudinal study has potential to identify predictors of adolescent substance misuse, particularly if it can incorporate a wide range of potential causal factors, both proximal and distal, and their influence on numerous social, psychological and biological mechanisms[4]. Here we apply machine learning to a wide range of data from a large sample of adolescents ($n = 692$) to generate models of current and future adolescent alcohol misuse that incorporate brain structure and function, individual personality and cognitive differences, environmental factors (including gestational cigarette and alcohol exposure), life experiences, and candidate genes. These models were accurate and generalized to novel data, and point to life experiences, neurobiological differences and personality as important antecedents of binge drinking. By identifying the vulnerability factors underlying individual differences in alcohol misuse, these models shed light on the aetiology of alcohol misuse and suggest targets for prevention.

Alcohol misuse is common among adolescents[5]: slightly over 40% of all 13–14-year-old adolescents in the USA report alcohol use and 10% of this age group exhibit regular use. These figures rise to almost 65% for any alcohol use and 27% who report regular use by age 16 years. This is of concern as murine models demonstrate that adolescents are more vulnerable to alcohol-induced neurotoxicity than adults[1]. Early alcohol use is a strong risk factor for adult alcohol dependence[6] and therefore identifying inter-individual vulnerabilities and predictors of alcohol use in human adolescents is of importance. Generating such predictors, however, is challenging, not least because large sample sizes are needed to provide accurate estimates of the small effect sizes that prevail in the biological sciences[7,8]. Therefore, previous prospective studies, which typically focus on just one type of risk factor, have necessarily yielded modest predictions of future alcohol misuse. Moreover, previous classification approaches incorporating biological data have often been flawed due to overfitting[9,10,11].

Personality measures, particularly those assessing traits conferring risk for substance misuse, can identify adolescents at high risk of substance misuse[12]. Life events in early adolescence, such as parental divorce[13], can also serve as predictors of future alcohol use. A number of candidate genes for alcohol dependence have been identified[14], although the overall risk conveyed by any one polymorphism is small[15]. Cognitive factors such as executive function (for example, inhibitory control), but not attention and visual memory, distinguished non-substance-using siblings of substance misusers from healthy controls[16]. Response inhibition was a modest predictor of adolescent alcohol misuse (explaining about 1% of variance) in a large sample of adolescents[17]. Until now, there have been no large-sample prospective studies examining the neural correlates of alcohol misuse, but there is some evidence of a reduction in brain activity during tests of inhibitory control for adolescents who subsequently engaged in heavy alcohol use[18].

Here, we construct models of current and future adolescent binge drinking by combining a wide range of data (Extended Data Table 1) from the IMAGEN project[19,20], a multi-dimensional longitudinal study of adolescent development, using regularized logistic regression[21] (Extended Data Fig. 1). First (Analysis 1), we identified the characteristics discriminating 115 14-year-old binge drinkers (a minimum of three lifetime binge drinking episodes leading to drunkenness by age 14) from 150 14-year-old controls (non-binge drinkers, a maximum of two lifetime uses of alcohol until at least the age of 16; see Extended Data Table 2 for participant details) returning an area-under-the-curve (AUC) receiver-operator characteristic (ROC) value of 0.96 (95% CI = 0.93–0.98; see Extended Data Table 3a for all beta weights). At the optimum point in the ROC curve, 91% of binge drinkers and 91% of non-binge drinkers were correctly classified, significantly better than chance ($P = 8.0 \times 10^{-61}$).

[1]Department of Psychiatry, University of Vermont, Burlington, Vermont 05401, USA. [2]Department of Psychology, University College Dublin, Dublin 4, Ireland. [3]Department of Radiology, University of Vermont, Burlington, Vermont 05401, USA. [4]Vermont Center for Children, Youth, and Families, University of Vermont, Burlington, Vermont 05401, USA. [5]Department of Pediatrics, University of Vermont, Burlington, Vermont 05401, USA. [6]Department of Psychology, University of Vermont, Burlington, Vermont 05401, USA. [7]Institut National de la Santé et de la Recherche Médicale, INSERM CEA Unit 1000 "Imaging & Psychiatry", University Paris Sud, 91400 Orsay, France. [8]Department of Psychiatry, Orsay Hospital, 4 place du General Leclerc, 91400 Orsay, France. [9]Department of Cognitive and Clinical Neuroscience, Central Institute of Mental Health, Medical Faculty Mannheim, Heidelberg University, 68159 Mannheim, Germany. [10]Institute of Psychiatry, King's College London, London SE5 8AF, UK. [11]Institute of Neuroscience, Trinity College Dublin, Dublin 2, Ireland. [12]Department of Systems Neuroscience, Universitätsklinikum Hamburg Eppendorf, 20246 Hamburg, Germany. [13]Department of Psychology, Stanford University, Stanford, California 94305, USA. [14]Department of Psychiatry, Université de Montreal, CHU Ste Justine Hospital, Montreal H3T 1C5, Canada. [15]Department of Addictive Behaviour and Addiction Medicine, Heidelberg University, 68159 Mannheim, Germany. [16]14 CEA, DSV, I2BM, Neurospin bat 145, 91191 Gif-Sur-Yvette, France. [17]Department of Psychiatry and Psychotherapy, Campus Charité Mitte, Charité–Universitätsmedizin Berlin 10117, Germany. [18]Department of Psychiatry and Neuroimaging Center, Technische Universität Dresden, 01062 Dresden, Germany. [19]School of Physics and Astronomy, University of Nottingham, Nottingham NG7 2RD, UK. [20]Physikalisch-Technische Bundesanstalt (PTB), 10587 Berlin, Germany. [21]School of Psychology, University of Nottingham, Nottingham NG7 2RD, UK. [22]AP-HP Department of Adolescent Psychopathology and Medicine, Maison de Solenn, University Paris Descartes, 75006 Paris, France. [23]Neuroscience Graduate Program, University of Vermont, Burlington, Vermont 05401, USA. [24]Rotman Research Institute, University of Toronto, Toronto, Ontario M5R 0A3, Canada. [25]Montreal Neurological Institute, McGill University, H3A 2B4, Canada. [26]The Hospital for Sick Children, University of Toronto, Toronto, Ontario M5G 0A4, Canada. [27]Behavioural and Clinical Neuroscience Institute and Department of Psychology, University of Cambridge, Cambridge CB2 1TN, UK. [28]MRC Social, Genetic and Developmental Psychiatry (SGDP) Centre, London, London WC2R 2LS, UK.
†A list of authors and affiliations appears at the end of the paper.

At the maximum *F*-score value, this classification accuracy corresponds to a precision rate of 87% (that is, those identified as binge drinkers who are actually binge drinkers) and a recall rate of 99% (that is, binge drinkers that are successfully detected; Extended Data Fig. 2a, b).

The model reported in Analysis 1, although highly accurate, was dominated by the inclusion of smoking, which often co-occurs with alcohol use. In Analysis 2, therefore, we removed smoking and re-ran the analyses (see Extended Data for all additional analyses with smoking included), which resulted in an AUC of 0.90 (95% CI = 0.86–0.93). At the optimum point in the ROC curve, 82% of binge drinkers and 89% of non-binge drinkers were correctly classified ($P = 8.8 \times 10^{-48}$). At the maximum *F*-score value the precision rate was 87% and the recall rate was 89% (Extended Data Fig. 2e, f). The features included in this model, and their strength of association with group membership, are displayed in Fig. 1a.

Figure 2a displays the brain regions that most consistently discriminated current binge drinkers from non-binge-drinkers (see Extended Data Fig. 3 for the contributions of each brain feature). The most robust brain classifiers were in ventromedial prefrontal cortex (vmPFC) and the left inferior frontal gyrus (IFG). The vmPFC grey matter volume was smaller in the current binge drinkers and this group, compared to controls, also showed decreased activity when anticipating or receiving a reward, but increased activity when processing angry faces. In the left IFG, current binge drinkers had smaller volumes and reduced activity when anticipating and receiving rewards and when processing angry faces.

The performance of each domain on its own (Analysis 3), both with and without age-14 smoking, is displayed in Extended Data Fig. 4a. The History and Personality domains were each accurate classifiers (AUC > 0.8). Next, we sought to quantify the unique contribution of each domain to the classification of current binge drinkers both with (Analysis 4) and without (Analysis 5) age-14 smoking. To this end, we iteratively removed each domain from the full model (re-calculating the optimum elastic net parameters), and observed the relative reduction in classification accuracy (Extended Data Fig. 4b, c). The History domain contributed the greatest unique variance to the model (significant correlations among features are displayed in Extended Data Fig. 5). The results of external generalizations of the current binge drinking models with and without nicotine (Analyses 6 and 7, respectively) are displayed in Extended Data Fig. 2c, d, g, h.

We have described the profile of current alcohol misusers while also demonstrating the efficacy of our modelling approach. However, to

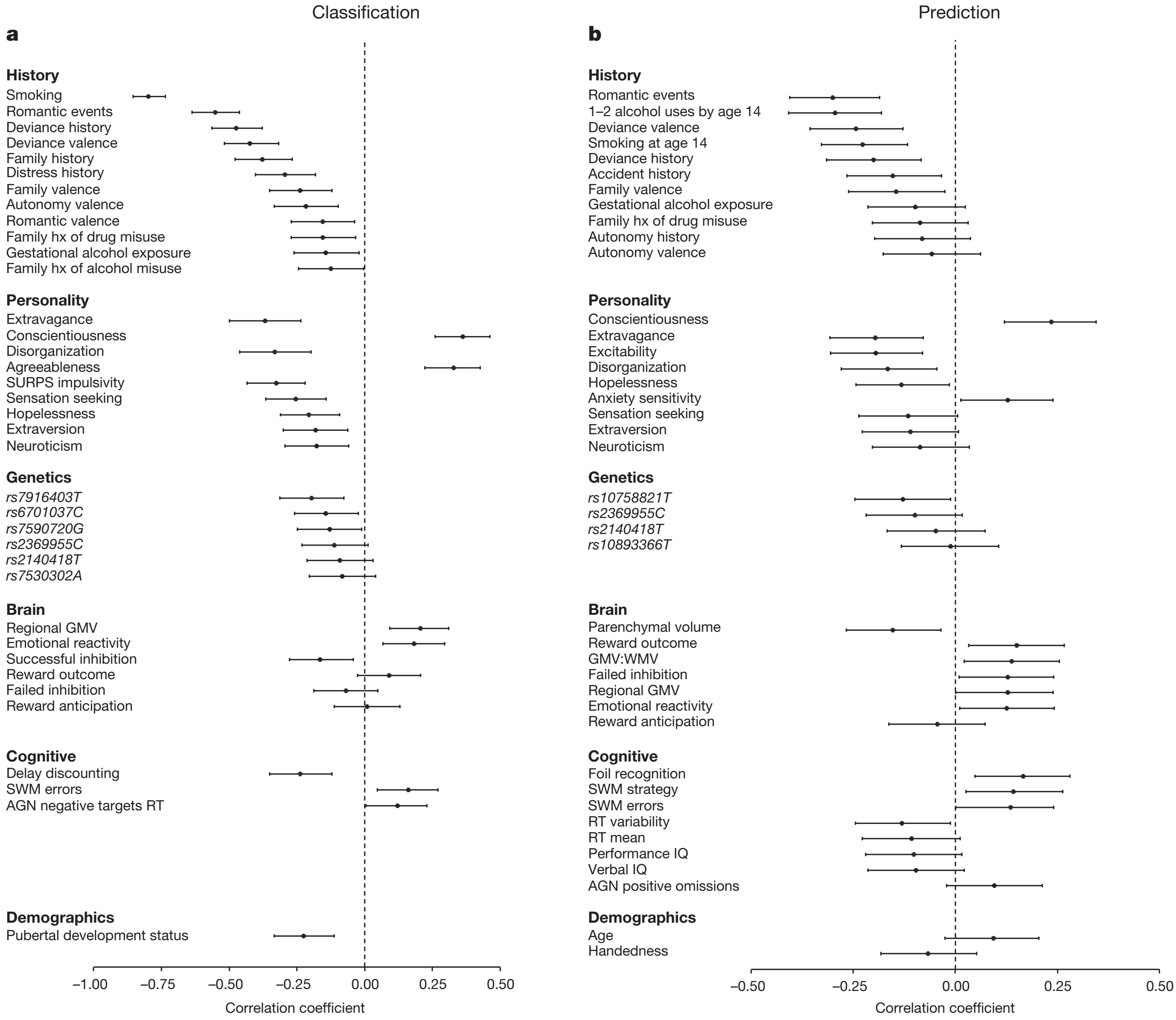


Figure 1 | The relationship between group membership and each feature that was present in at least 9 folds of the final model. Position on the horizontal represents the point-biserial correlation statistic (*r*) between each feature and group membership. Negative *r* values indicate that higher scores are associated with an increased likelihood to engage in binge drinking at 14. Error bars represent 95% confidence intervals (calculated using 10,000 bootstraps). **a**, Analyses 1 and 2, the classification of binge drinking at age 14 years (*n* = 265). **b**, Analysis 8 predicting binge drinking at age 16 years (*n* = 271). AGN, affective go/no go; hx, history; SURPS, substance use risk profile scale; SWM, spatial working memory; GMV, grey matter volume; WMV, white matter volume.

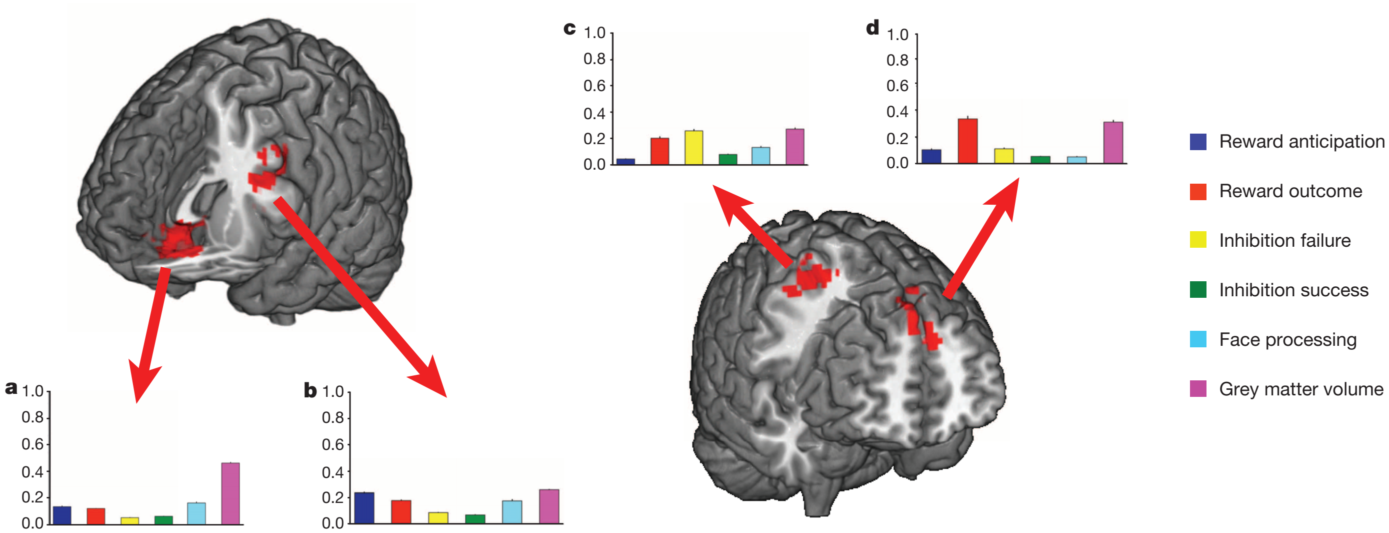


Figure 2 | Brain regions associated with binge drinking and the relative contribution of each brain metric to the classification. The average beta weight for each brain metric (normalized to sum to 1 and averaged over the ten outer folds). Error bars depict standard errors of the mean across the folds. **a**, **b**, Brain regions that classify binge drinking at age 14, Analyses 1 and 2 ($n = 265$). The most robust brain classifiers were in ventromedial prefrontal cortex (**a**) and the left inferior frontal gyrus (**b**). **c**, **d**, Brain regions that predict binge drinking at age 16, Analysis 8 ($n = 271$). The most robust brain predictors of future binge drinking were the right precentral gyrus (**c**) and bilateral superior frontal gyrus (**d**).

identify risk factors for adolescent alcohol misuse, a matter of clinical relevance, a model that predicts future binge drinking is required. Thus, in Analysis 8, we compared 121 future binge drinkers (a maximum of two drink occasions by age 14 and a minimum of three lifetime binge drinking episodes by age 16) to the 150 controls described previously. This model had an AUC of 0.75 (95% CI = 0.69–0.80; Extended Data Fig. 2i, j). At the optimum point in the AUC curve, 73% of non-binge drinkers and 66% of binge drinkers were correctly classified, significantly better than chance ($P = 4.2 \times 10^{-17}$) given a base rate of 45% binge drinkers. This corresponds to a precision rate of 64% and a recall rate of 93% at the maximum *F*-score value. The features of the final model are displayed in Fig. 1b. Figure 2b displays the brain regions that discriminated future binge drinkers from non-binge-drinkers and the contributions of each functional/structural feature are displayed in Extended Data Fig. 6.

Next, we examined each domain on its own (Analysis 9). History was still the most predictive domain; however, now its influence was broadly comparable to Brain and Personality (Extended Data Fig. 4d), although the unique contribution of History was more apparent when each domain was iteratively removed from the model (Analysis 10; Extended Data Fig. 4e). Significant correlations among the features are displayed in Extended Data Fig. 7.

Our profile of adolescent binge drinking used a large sample and was internally valid, in that it generalized well using cross-validation. However, an outstanding question is whether or not this profile would be applicable to a new sample with different levels of alcohol consumption, which would speak to the dimensional nature of substance misuse[22]. Thus, we applied the prediction model from Analysis 8 to a new sample from the IMAGEN study (Analysis 11): all subjects had between 3–5 lifetime drink occasions (that is, a score of 2 on the substance misuse questionnaire) but no binge drinking episodes by age 14; by age 16, 61 of these still had no binge-drinking episodes whereas 55 participants had at least 3 binge-drinking episodes. Application of the model (without age-14 drinking as this was the same for all participants) resulted in similar predictability to that reported above: ROC AUC = 0.75 (95% CI = 0.66–0.83). At the optimal point of the AUC 77% of binge drinkers and 67% of non-binge-drinkers were correctly assigned ($P = 2.71 \times 10^{-8}$). At the maximum *F*-score value, this corresponds to a precision rate of 65% and a recall rate of 93%. The most robust brain predictors of future binge drinking were the right middle and precentral gyri (Brodmann Area 6) and bilateral superior frontal gyrus (Brodmann Area 9). At age 14 future binge drinkers had reduced grey matter volume but increased activity when receiving a reward in the superior frontal gyrus compared to controls. In premotor cortex, future binge drinkers showed greater grey matter volume and greater activity when failing to inhibit.

A number of features were common to both current and future alcohol misuse (Analyses 2 and 8). Life events, such as a romantic or sexual relationship, were strong classifiers for both current and future binge drinkers. Personality measures associated with binge drinking included the novelty-seeking trait from the temperament and character inventory (TCI) psychobiological model of personality[23]. This trait identifies the behaviour of searching for, and feeling rewarded by, novel experiences and is regarded as a heritable, dopamine-related temperament: higher scores on Disorderliness and Extravagance (a tendency to approach reward cues) characterized both current and future binge drinkers. Conscientiousness (the degree to which an individual is organized, controlled and motivated to achieve a desired goal) was lower in both current and future binge drinkers.

Some features differed in their utility to classify current and future binge drinkers. Disruptive family events, the personality trait of agreeableness, more developed pubertal status, impulsivity and higher delay discounting (the tendency to devalue future rewards) classified current, but not future, binge drinkers. In contrast, the anxiety sensitivity subscale of the substance use risk profile scale (SURPS)[24] (fear of anxiety-related emotions and sensations due to beliefs that these emotions and sensations could lead to harmful consequences) predicted non-binge drinking at age 16, not at age 14. Parenchymal volume and grey:white matter ratio predicted future, but not current binge drinking. The most prominent brain regions for classifying current binge drinkers included the vmPFC and the left lateral PFC, areas that have been implicated in emotional regulation of bingeing behaviour[25,26]. Whereas emotional processing areas were implicated in age-14 binge drinking, predicting age-16 binge drinkers from data at age 14 relied relatively more on regions associated with failed inhibitory control and reward outcome and on local and global brain structure. Notably, even 1–2 lifetime alcohol occasions by age 14 was sufficient to be an important predictor of future binge drinking at age 16.

We have identified a generalizable risk profile for alcohol misuse initiation. In contrast with the classification of current binge drinkers, which was primarily a function of the History domain, the prediction of future binge drinking relied relatively more on a combination of three domains: History, Personality and Brain (individual ROC AUCs of 0.68, 0.67 and 0.63, respectively; Analysis 9). Thus, these results point to the value of a multi-domain analysis for predicting adolescent alcohol misuse and speak to the multiple causal factors for alcohol misuse. Further, we note that the influence of any one feature in isolation was modest, consistent with data showing that effect sizes in previous studies with smaller samples are likely to have been overestimated[7,9]. Given that the odds of adult alcohol dependence can be reduced by 10% for each year drinking onset is delayed in adolescence[27], this risk profile may facilitate the development of targeted interventions[24,28], which often yield higher effect sizes than general approaches[29].

METHODS SUMMARY

Informed consent was obtained from all subjects and their parents/guardians. We collected a wide range of data at age 14, which were arranged into the following domains: Brain, Personality, Cognition, History, Genetics and Demographics (Extended Data Table 1). Substance misuse data were acquired at both ages 14 and 16. Functional brain activity was recorded during reward anticipation and outcome, successful and unsuccessful inhibitions on a test of motor inhibitory control, and a test of emotional reactivity to angry faces. Structural brain data consisted of regional grey matter volume, total parenchymal volume, and white:grey matter ratio. Personality data included both broad personality traits and those specifically related to substance misuse. Cognitive measures assessed IQ, delay discounting, spatial working memory, attentional biases for affective stimuli and behavioural measures from the functional imaging tasks. The History domain included life events, family history of alcohol and drug misuse and gestational alcohol and cigarette exposure. We assessed 15 candidate genes related to alcohol abuse[14], and demographic features included sex, pubertal development status and socioeconomic status. To construct the models, a multistep procedure was used to create summary scores from brain images first, which were then combined with the other data (Extended Data Fig. 1). Classification was conducted using logistic regression with elastic net regularization[21], allowing the inclusion of correlated features in sparse model fits (that is, potentially selecting a subset of features). We report data from the test sets of a tenfold cross validation.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 6 June 2013; accepted 23 April 2014.

Published online 2 July; corrected online 13 August 2014 (see full-text HTML version for details).

1. Crews, F. T., Braun, C. J., Hoplight, B., Switzer, R. C. III & Knapp, D. J. Binge ethanol consumption causes differential brain damage in young adolescent rats compared with adult rats. *Alcohol. Clin. Exp. Res.* **24,** 1712–1723 (2000).
2. Ersche, K. D. *et al.* Abnormal brain structure implicated in stimulant drug addiction. *Science* **335,** 601–604 (2012).
3. Volkow, N. D. *et al.* High levels of dopamine D2 receptors in unaffected members of alcoholic families: possible protective factors. *Arch. Gen. Psychiatry* **63,** 999–1008 (2006).
4. Cloninger, C. R. Neurogenetic adaptive mechanisms in alcoholism. *Science* **236,** 410–416 (1987).
5. Swendsen, J. *et al.* Use and abuse of alcohol and illicit drugs in US adolescents: results of the National Comorbidity Survey-Adolescent Supplement. *Arch. Gen. Psychiatry* **69,** 390–398 (2012).
6. Grant, J. D. *et al.* Adolescent alcohol use is a risk factor for adult alcohol and drug dependence: evidence from a twin design. *Psychol. Med.* **36,** 109–118 (2006).
7. Ioannidis, J. P. Why most published research findings are false. *PLoS Med.* **2,** e124 (2005).
8. Button, K. S. *et al.* Power failure: why small sample size undermines the reliability of neuroscience. *Nature Rev. Neurosci.* **14,** 365–376 (2013).
9. Whelan, R. & Garavan, H. Prediction inflation in neuroimaging. *Biol. Psychiatry* **75,** 746–748 (2014).
10. Bellazzi, R. & Zupan, B. Predictive data mining in clinical medicine: current issues and guidelines. *Int. J. Med. Inform.* **77,** 81–97 (2008).
11. Ambroise, C. & McLachlan, G. Selection bias in gene extraction on the basis of microarray gene-expression data. *Proc. Natl Acad. Sci. USA* **99,** 6562–6566 (2002).
12. Castellanos-Ryan, N., O'Leary-Barrett, M., Sully, L. & Conrod, P. Sensitivity and specificity of a brief personality screening instrument in predicting future substance use, emotional, and behavioral problems: 18-month predictive validity of the Substance Use Risk Profile Scale. *Alcohol. Clin. Exp. Res.* **37,** E281–E290 (2013).
13. Dube, S. *et al.* Adverse childhood experiences and the association with ever using alcohol and initiating alcohol use during adolescence. *J. Adolesc. Health* **38,** 444.e1–444.e10 (2006).
14. Rietschel, M. & Treutlein, J. The genetics of alcohol dependence. *Ann. NY Acad. Sci.* **1282,** 39–70 (2013).
15. Tyndale, R. F. Genetics of alcohol and tobacco use in humans. *Ann. Med.* **39,** 94–121 (2003).
16. Ersche, K. D. *et al.* Cognitive dysfunction and anxious-impulsive personality traits are endophenotypes for drug dependence. *Am. J. Psychiatry* **169,** 926–936 (2012).
17. Nigg, J. *et al.* Poor response inhibition as a predictor of problem drinking and illicit drug use in adolescents at risk for alcoholism and other substance use disorders. *J. Am. Acad. Child Adolesc. Psychiatry* **45,** 468–475 (2006).
18. Norman, A. L. *et al.* Neural activation during inhibition predicts initiation of substance use in adolescence. *Drug Alcohol Depend.* **119,** 216–223 (2011).
19. Schumann, G. *et al.* The IMAGEN study: reinforcement-related behaviour in normal brain function and psychopathology. *Mol. Psychiatry* **15,** 1128–1139 (2010).
20. Whelan, R. *et al.* Adolescent impulsivity phenotypes characterized by distinct brain networks. *Nature Neurosci.* **15,** 920–925 (2012).
21. Zou, H. & Hastie, T. Regularization and variable selection via the elastic net. *J. Roy. Stat. Soc. B* **67,** 301–320 (2005).
22. Robbins, T. W., Gillan, C., Smith, D., de Wit, S. & Ersche, K. Neurocognitive endophenotypes of impulsivity and compulsivity: towards dimensional psychiatry. *Trends Cogn. Sci.* **16,** 81–91 (2012).
23. Cloninger, C. R., Svrakic, D. M. & Przybeck, T. R. A psychobiological model of temperament and character. *Arch. Gen. Psychiatry* **50,** 975–990 (1993).
24. Conrod, P. J., Castellanos, N. & Mackie, C. Personality-targeted interventions delay the growth of adolescent drinking and binge drinking. *J. Child Psychol. Psychiatry* **49,** 181–190 (2008).
25. Goldstein, R. Z. & Volkow, N. Dysfunction of the prefrontal cortex in addiction: neuroimaging findings and clinical implications. *Nature Rev. Neurosci.* **12,** 652–669 (2011).
26. Hare, T. A., Camerer, C. F. & Rangel, A. Self-control in decision-making involves modulation of the vmPFC valuation system. *Science* **324,** 646–648 (2009).
27. Grant, B. F., Stinson, F. S. & Harford, T. C. Age at onset of alcohol use and DSM-IV alcohol abuse and dependence: a 12-year follow-up. *J. Subst. Abuse* **13,** 493–504 (2001).
28. Ghahremani, D. G. *et al.* Effects of the Youth Empowerment Seminar on impulsive behavior in adolescents. *J. Adolesc. Health* **53,** 139–141 (2013).
29. Gottfredson, D. C. & Wilson, D. B. Characteristics of effective school-based substance abuse prevention. *Prev. Sci.* **4,** 27–38 (2003).

Supplementary Information is available in the online version of the paper.

Acknowledgements This work received support from the following sources: the European Union-funded FP6 Integrated Project IMAGEN (Reinforcement-related behaviour in normal brain function and psychopathology) (LSHM-CT- 2007-037286), the FP7 projects IMAGEMEND (602450; IMAging GEnetics for MENtal Disorders) and MATRICS (603016), the Innovative Medicine Initiative Project EU-AIMS (115300-2), a Medical Research Council Programme Grant ''Developmental pathways into adolescent substance abuse'' (93558), the Swedish funding agency FORMAS, the Medical Research Council and the Wellcome Trust (Behavioural and Clinical Neuroscience Institute, University of Cambridge), the National Institute for Health Research (NIHR) Biomedical Research Centre at South London and Maudsley NHS Foundation Trust and King's College London, the Bundesministerium für Bildung und Forschung (BMBF grants 01GS08152; 01EV0711; eMED SysAlc 01ZX1311A; Forschungsnetz AERIAL), the Deutsche Forschungsgemeinschaft (DFG): Reinhart-Koselleck Award SP 383/5-1 and grants SM 80/7-1, SFB 940/1, FOR 1617), the French MILDT (Mission Interministérielle de Lutte contre la Drogue et la Toxicomanie), the CENIR (Centre de NeuroImagerie de Recherche, Pr. S. Lehéricy) within the ICM institute, the National Institute of Mental Health (MH082116), a National Institutes of Health Center of Biomedical Research Excellence award P20GM103644 from the National Institute of General Medical Sciences and the Tobacco Centers of Regulatory Science award P50DA036114. The authors acknowledge the Vermont Advanced Computing Core which is supported by NASA (NNX 06AC88G), at the University of Vermont for providing high performance computing resources that have contributed to the research results reported within this paper.

Author Contributions T.B., G.J.B., A.L.W.B., C.B., F.M.C., P.J.C., H.F., M.F.-B., J.G., H.G., P.G., A.H., B.I., K.M., J.-L.M., F.N., T.P., M.R., C.L., Z.P., M.-L.P.-M., M.N.S., A.S., M.R. and T.W.R. acquired the data. R. Whelan., H.G., C.A.O. and N.O. analysed the behavioural data. G.G. calculated the family history data. R. Whelan, R. Watts and E.A. carried out neuroimaging data processing and analysis. R.R.A., V.F. and G.S. carried out genotyping and genetic analysis. R. Whelan and H.G. prepared the manuscript. C.A.O, P.J.C., J.G., T.P., T.W.R. and G.S. edited the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to R. Whelan (Robert.whelan@ucd.ie) or H.G. (Hugh.garavan@uvm.edu).

The IMAGEN Consortium

Lisa Albrecht[1], Mercedes Arroyo[2], Semiha Aydin[3],Christine Bach[4], Alexis Barbot[5], Zuleima Bricaud[6], Uli Bromberg[7], Ruediger Bruehl[3], Anna Cattrell[8], Katharina Czech[1], Jeffrey Dalley[2], Sylvane Desrivieres[8], Tahmine Fadai[7], Birgit Fuchs[9], Fanny Gollier Briand[6], Kay Head[10], Bert Heinrichs[11], Nadja Heym[10], Thomas Hübner[12], Albrecht Ihlenfeld[3], James Ireland[13], Nikolay Ivanov[1], Tianye Jia[8], Jennifer Jones[14], Agnes

Kepa[8], Dirk Lanzerath[11], Mark Lathrop[15], Hervé Lemaitre[6], Katharina Lüdemann[1], Lourdes Martinez-Medina[8], Xavier Mignon[16], Ruben Miranda[6], Kathrin Müller[12], Charlotte Nymberg[8], Jani Pentilla[6], Jean-Baptiste Poline[5], Luise Poustka[4], Michael Rapp[1], Stephan Ripke[12], Sarah Rodehacke[12], John Rogers[13], Alexander Romanowski[1], Barbara Ruggeri[8], Christine Schmäl[4], Dirk Schmidt[12], Sophia Schneider[7], Markus Schroeder[17], Florian Schubert[3], Wolfgang Sommer[4], Rainer Spanagel[4], David Stacey[8], Sabina Steiner[4], Dai Stephens[18], Nicole Strache[1], Maren Struve[4], Amir Tahmasebi[19], Lauren Topper[8], Helene Vulser[6], Bernadeta Walaszek[3], Helen Werts[8], Steve Williams[8], C. Peng Wong[8], Juliana Yacubian[7] & Veronika Ziesch[12].

[1]Campus Charité Mitte, Charité–Universitätsmedizin, Berlin 10117, Germany. [2]University of Cambridge, Cambridge CB2 1TN, UK. [3]Physikalisch-Technische Bundesanstalt (PTB), 10587 Berlin, Germany. [4]Central Institute of Mental Health, Medical Faculty Mannheim, Heidelberg University, 68159 Mannheim, Germany. [5]Commissariat à l'Energie Atomique, 14 CEA, DSV, I2BM, Neurospin bat 145, 91191 Gif-Sur-Yvette, France. [6]Institut National de la Santé et de la Recherche Médicale, INSERM CEA Unit 1000 ''Imaging & Psychiatry'', University Paris Sud, 91400 Orsay, France. [7]Universitätsklinikum Hamburg Eppendorf, 20246 Hamburg, Germany. [8]Institute of Psychiatry, King's College London, London SE5 8AF, UK. [9]GABO:milliarium mbH & Co. KG 80333 Munich, Germany. [10]University of Nottingham, Nottingham NG7 2RD, UK. [11]Deutsches Referenzzentrum für Ethik, D 53113 Bonn, Germany. [12]Technische Universität Dresden, 01062 Dresden, Germany. [13]Delosis, Twickenham, Middlesex TW1 4AE, UK. [14]Trinity College Dublin, Dublin 2, Ireland. [15]Centre National de Génotypage, 91057 Evry Cedex, France. [16]PERTIMM, 92600 Asnières-Sur-Seine, France. [17]Tembit Software GmbH, 13507 Berlin, Germany. [18]University of Sussex, Brighton BN1 9RH, UK. [19]University of Toronto, Toronto, Ontario M5G 0A4, Canada.

doi:10.1038/nature13425

A common Greenlandic *TBC1D4* variant confers muscle insulin resistance and type 2 diabetes

Ida Moltke[1,2]*, Niels Grarup[3]*, Marit E. Jørgensen[4], Peter Bjerregaard[5], Jonas T. Treebak[6], Matteo Fumagalli[7], Thorfinn S. Korneliussen[8], Marianne A. Andersen[6], Thomas S. Nielsen[6], Nikolaj T. Krarup[3], Anette P. Gjesing[3], Juleen R. Zierath[6,9], Allan Linneberg[10], Xueli Wu[11], Guangqing Sun[11], Xin Jin[11], Jumana Al-Aama[11,12], Jun Wang[3,11,12,13,14], Knut Borch-Johnsen[15], Oluf Pedersen[3], Rasmus Nielsen[7,16], Anders Albrechtsen[1] & Torben Hansen[3,17]

The Greenlandic population, a small and historically isolated founder population comprising about 57,000 inhabitants, has experienced a dramatic increase in type 2 diabetes (T2D) prevalence during the past 25 years[1]. Motivated by this, we performed association mapping of T2D-related quantitative traits in up to 2,575 Greenlandic individuals without known diabetes. Using array-based genotyping and exome sequencing, we discovered a nonsense p.Arg684Ter variant (in which arginine is replaced by a termination codon) in the gene *TBC1D4* with an allele frequency of 17%. Here we show that homozygous carriers of this variant have markedly higher concentrations of plasma glucose ($\beta = 3.8$ mmol l^{-1}, $P = 2.5 \times 10^{-35}$) and serum insulin ($\beta = 165$ pmol l^{-1}, $P = 1.5 \times 10^{-20}$) 2 hours after an oral glucose load compared with individuals with other genotypes (both non-carriers and heterozygous carriers). Furthermore, homozygous carriers have marginally lower concentrations of fasting plasma glucose ($\beta = -0.18$ mmol l^{-1}, $P = 1.1 \times 10^{-6}$) and fasting serum insulin ($\beta = -8.3$ pmol l^{-1}, $P = 0.0014$), and their T2D risk is markedly increased (odds ratio (OR) = 10.3, $P = 1.6 \times 10^{-24}$). Heterozygous carriers have a moderately higher plasma glucose concentration 2 hours after an oral glucose load than non-carriers ($\beta = 0.43$ mmol l^{-1}, $P = 5.3 \times 10^{-5}$). Analyses of skeletal muscle biopsies showed lower messenger RNA and protein levels of the long isoform of TBC1D4, and lower muscle protein levels of the glucose transporter GLUT4, with increasing number of p.Arg684Ter alleles. These findings are concomitant with a severely decreased insulin-stimulated glucose uptake in muscle, leading to postprandial hyperglycaemia, impaired glucose tolerance and T2D. The observed effect sizes are several times larger than any previous findings in large-scale genome-wide association studies of these traits[2–4] and constitute further proof of the value of conducting genetic association studies outside the traditional setting of large homogeneous populations.

Genetic association studies have traditionally been performed in large homogeneous populations. However, several studies have shown that it can be valuable to use founder populations[5], and there are similar advantages to using small and historically isolated populations. These advantages include increased statistical power to detect associations, owing to extended linkage disequilibrium and to an increased probability that deleterious variants overcome their selective disadvantage and reach high allele frequencies as a result of substantial genetic drift over many generations. Therefore, we aimed at identifying genetic variants associated with glucose homeostasis in the Greenlandic population, which is a small and historically isolated founder population, by association mapping of four T2D-related traits: plasma glucose and serum insulin levels at fasting and 2 h after an oral glucose load.

We successfully genotyped 2,733 participants in the Inuit Health in Transition (IHIT) cohort[6], sampled from 12 regions in Greenland (Fig. 1a), with the Illumina Cardio-Metabochip[7] (Metabochip) (Extended Data Table 1). We observed a high degree of linkage disequilibrium compared with Europeans (Extended Data Fig. 1a) and a high degree of European admixture (Fig. 1b). Additionally, as a natural consequence of the fact that the cohort constitutes almost 5% of the population, we identified more than 1,000 close relationships (siblings or parent–offspring) (Extended Data Fig. 1b). Population structure can lead to both decreased statistical power and increased type II error rates[8]. To avoid the latter, we therefore performed association analyses using a linear mixed model, which takes both admixture and relatedness into account. Genomic control inflation factors showed no inflation (range, 0.937–0.995; Fig. 2a and Extended Data Fig. 2).

Discovery analyses were performed for up to 2,575 IHIT participants without previously known T2D using an additive model. We found that the minor allele of rs7330796 was strongly associated with a higher 2-h plasma glucose level ($P = 4.2 \times 10^{-17}$) and a higher 2-h serum insulin level ($P = 6.4 \times 10^{-10}$) (Fig. 2a, b and Extended Data Fig. 2). These associations were replicated in the B99 Greenlandic cohort[9] ($P = 4.2 \times 10^{-6}$ for 2-h plasma glucose and $P = 2.7 \times 10^{-5}$ for 2-h serum insulin).

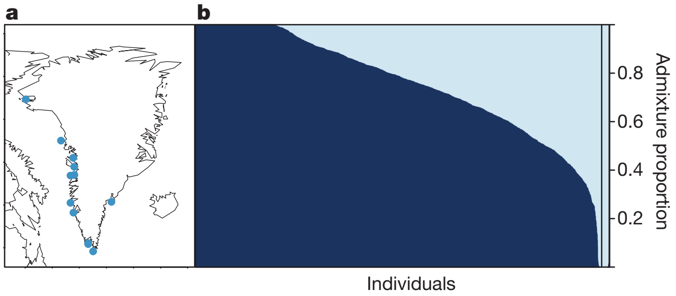


Figure 1 | Greenlandic study population. **a**, Sampling locations in Greenland. **b**, Estimated admixture proportions of Inuit and European ancestry. The admixture proportions were estimated assuming two source populations ($K = 2$). The estimates are both for the 2,733 individuals in the Greenlandic sample (IHIT), depicted to the left of the vertical line, and for 50 Danes, to the right of the vertical line.

[1] The Bioinformatics Centre, Department of Biology, University of Copenhagen, 2200 Copenhagen, Denmark. [2] Department of Human Genetics, University of Chicago, Chicago, Illinois 60637, USA. [3] The Novo Nordisk Foundation Center for Basic Metabolic Research, Section of Metabolic Genetics, Faculty of Health and Medical Sciences, University of Copenhagen, 2100 Copenhagen, Denmark. [4] Steno Diabetes Center, 2820 Gentofte, Denmark. [5] National Institute of Public Health, University of Southern Denmark, 1353 Copenhagen, Denmark. [6] The Novo Nordisk Foundation Center for Basic Metabolic Research, Section of Integrative Physiology, Faculty of Health and Medical Sciences, University of Copenhagen, 2200 Copenhagen, Denmark. [7] Department of Integrative Biology, University of California, Berkeley, California 94720, USA. [8] Centre for GeoGenetics, Natural History Museum of Denmark, University of Copenhagen, 1350 Copenhagen, Denmark. [9] Department of Molecular Medicine and Surgery, Karolinska Institute, 171 77 Stockholm, Sweden. [10] Research Centre for Prevention and Health, Glostrup University Hospital, 2600 Glostrup, Denmark. [11] BGI-Shenzhen, Shenzhen 518083, China. [12] The Department of Genetic Medicine, Faculty of Medicine and Princess Al Jawhara Albrahim Center of Excellence in the Research of Hereditary Disorders, King Abdulaziz University, Jeddah 21589, Saudi Arabia. [13] Department of Biology, University of Copenhagen, 2200 Copenhagen, Denmark. [14] Macau University of Science and Technology, Macau 999078, China. [15] Holbaek Hospital, 4300 Holbaek, Denmark. [16] Department of Statistics, University of California, Berkeley, California 94720, USA. [17] Faculty of Health Sciences, University of Southern Denmark, 5000 Odense, Denmark.
*These authors contributed equally to this work.

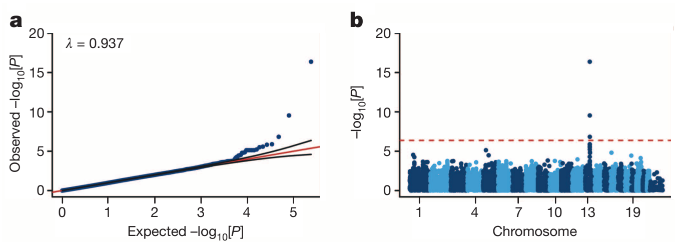


Figure 2 | Associations between 2-h plasma glucose levels and genotypes, as determined by Metabochip assay. Tests were performed using an additive linear mixed model in 2,540 individuals from the IHIT study who were not known to have T2D and for whom valid 2-h plasma glucose data were available. **a**, A quantile–quantile (QQ) plot of the observed $-\log_{10}[P]$ values (y axis) versus the $-\log_{10}[P]$ values expected under the null hypothesis of no association (x axis). The red line shows $x = y$, and the black lines demarcate the 95% confidence interval. The λ value is the genomic control inflation factor. **b**, A Manhattan plot of the observed $-\log_{10}[P]$ values. The dashed horizontal line indicates a 0.05 significance threshold after Bonferroni correction for multiple testing. The lowest P value is for rs7330796 on chromosome 13.

The rs7330796 variant was selected for inclusion on the Metabochip because it was in the top 5,000 candidate single nucleotide polymorphisms (SNPs) for association with waist-to-hip ratio[7] and it has not previously been reported to be associated with any of the four examined T2D-related traits or with T2D. The variant is located in intron 11 of *TBC1D4* and is neither in high linkage disequilibrium with neighbouring variants on the Metabochip nor situated inside a long range linkage disequilibrium block (Extended Data Fig. 3a, b). To locate the causal variation in the region, we performed exome sequencing of nine trios, and we identified four coding SNPs in high linkage disequilibrium ($r^2 > 0.8$) with rs7330796 (Extended Data Table 2). We genotyped these SNPs and found that p.Arg684Ter, a nonsense polymorphism in *TBC1D4* (c.2050C>T, rs61736969), was strongly associated with 2-h plasma glucose levels ($P = 3.6 \times 10^{-25}$) in the IHIT cohort (Table 1 and Fig. 3a). Conditional analyses demonstrated that p.Arg684Ter was significantly associated with 2-h plasma glucose and 2-h serum insulin levels when conditioning on rs7330796 ($P = 1.3 \times 10^{-9}$ and $P = 8.9 \times 10^{-9}$, respectively), whereas rs7330796 was not associated with the two traits when conditioning on p.Arg684Ter ($P = 0.47$ and $P = 0.09$) (Fig. 3a). Additionally, the mean 2-h plasma glucose levels for individuals with two copies of the minor rs7330796 allele increased with increasing Inuit admixture proportion (Extended Data Fig. 4a), which is expected if a variant is not causative and if the linkage disequilibrium patterns differ between Inuit and Europeans. By contrast, the same was not true for p.Arg684Ter (Extended Data Fig. 4b). These findings suggest that p.Arg684Ter is the causative variant.

The mean 2-h plasma glucose levels stratified by p.Arg684Ter genotype suggested that the variant mainly has an effect in homozygous carriers, indicating a recessive inheritance (Fig. 3b). We therefore also performed analyses using a recessive model: that is, we compared homozygous

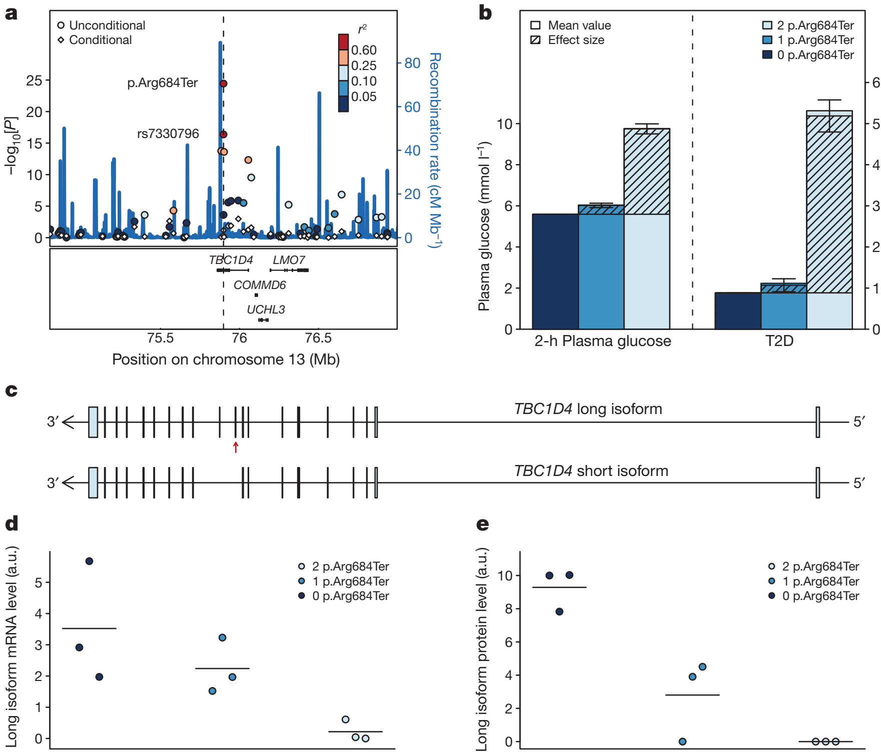


Figure 3 | Effect of the p.Arg684Ter nonsense polymorphism in *TBC1D4*. **a**, Association test results for all tested Metabochip SNPs in a 2-megabase (Mb) region around the p.Arg684Ter polymorphism (shown as a dashed vertical line). Each SNP is represented by a coloured circle. The position of the circle on the x axis shows the genomic position of the SNP. The position of the circle on the left y axis shows the $-\log_{10}[P]$ value of the SNP when testing for association with 2-h plasma glucose levels, as determined using an additive model. The colour of the circle indicates the extent of correlation (r^2) between the SNP and p.Arg684Ter. The circles representing p.Arg684Ter and rs7330796 are labelled (to the left of the circles). For every SNP, except for p.Arg684Ter, there is also a white diamond, which illustrates the P value obtained by testing for association conditional on p.Arg684Ter. The solid blue line illustrates the recombination rate from the Chinese HapMap (CHB) panel (in centimorgan (cM) per Mb, right y axis). The protein-coding genes in the genetic region are shown below the plot. **b**, The mean 2-h plasma glucose and the frequency of T2D for three genotypes (zero, one or two p.Arg684Ter alleles). Superimposed are the estimated effect sizes from the mixed model ± s.e.m. **c**, The two predominant isoforms of the *TBC1D4* gene illustrating which exons are transcribed: the long isoform, which has two additional exons (top), and the short isoform (bottom). Exons are depicted as boxes, and the location of the p.Arg684Ter polymorphism is indicated by a red arrow. **d**, The mRNA expression level of the long *TBC1D4* isoform in skeletal muscle from nine Greenlandic individuals (measured in arbitrary units (a.u.)). The mean value for each genotype group is shown as a horizontal line. **e**, Abundance of the long TBC1D4 protein isoform in skeletal muscle from nine individuals as quantified from western blot (measured in a.u.). The mean value for each genotype group is shown as a horizontal line.

Table 1 | Association of *TBC1D4* p.Arg684Ter with metabolic traits in the IHIT cohort

Trait	n	Additive model			Recessive model		
		$\beta_{\text{s.d.}}$ (95% CI)	β	P	$\beta_{\text{s.d.}}$ (95% CI)	β	P
Fasting plasma glucose (mmol l^{-1})	2,546	−0.13 (−0.2 to −0.064)	−0.048	0.00011	−0.45 (−0.63 to −0.27)	−0.18	1.1×10^{-6}
2-h Plasma glucose (mmol l^{-1})	2,511	0.35 (0.29 to 0.42)	1.1	**3.6×10^{-25}**	1.2 (0.99 to 1.4)	3.8	**2.5×10^{-35}**
Fasting serum insulin (pmol l^{-1})	2,546	−0.14 (−0.21 to −0.07)	−2.3	0.00012	−0.33 (−0.53 to −0.13)	−8.3	0.0014
2-h Serum insulin (pmol l^{-1})	2,511	0.29 (0.22 to 0.36)	57	**6.7×10^{-17}**	0.90 (0.71 to 1.1)	160	**1.5×10^{-20}**
Fasting serum C-peptide (pmol l^{-1})	2,546	−0.12 (−0.19 to −0.049)	−28	0.00092	−0.32 (−0.52 to −0.13)	−85	0.0012
2-h Serum C-peptide (pmol l^{-1})	2,511	0.30 (0.24 to 0.36)	360	**4.4×10^{-20}**	0.82 (0.65 to 1)	1,000	**8.5×10^{-20}**
HbA$_{1C}$ (%)	2,692	0.015 (−0.047 to 0.078)	0.015	0.63	0.2 (0.036 to 0.37)	0.10	0.017
HOMA-IR (mmol l^{-1} × pmol l^{-1})	2,546	−0.15 (−0.22 to −0.076)	−0.078	6.4×10^{-5}	−0.36 (−0.56 to −0.16)	−0.37	0.00047
ISI$_{0,120}$	2,487	−0.32 (−0.39 to −0.26)	−0.47	**1.4×10^{-20}**	−1.0 (−1.2 to −0.86)	−1.4	**1.6×10^{-27}**
HOMA-B (%)	2,545	−0.085 (−0.15 to −0.018)	−2.3	0.013	−0.12 (−0.31 to 0.062)	−4.2	0.19
T2D (cases/controls)	220/1,810	0.083 (0.059 to 0.11)	0.083	**2.1×10^{-11}**	0.37 (0.3 to 0.44)	0.37	**1.6×10^{-24}**
Fasting serum HDL-cholesterol (mmol l^{-1})	2,702	0.032 (−0.035 to 0.099)	0.019	0.34	0.098 (−0.082 to 0.28)	0.064	0.29
Fasting serum total cholesterol (mmol l^{-1})	2,566	−0.013 (−0.081 to 0.056)	−0.018	0.71	0.25 (0.064 to 0.44)	0.30	0.0086
Fasting serum triglyceride (mmol l^{-1})	2,702	−0.040 (−0.11 to 0.032)	−0.020	0.27	0.022 (−0.17 to 0.22)	0.038	0.82
BMI (kg m^{-2})	2,673	−0.036 (−0.11 to 0.036)	−0.19	0.32	0.047 (−0.15 to 0.24)	0.25	0.63

Results are shown for an additive and a recessive genetic model. For each trait, n is the number of individuals with genotype data for the specific variant and phenotype data for the specific trait. $\beta_{\text{s.d.}}$ is the effect size estimated using quantile-transformed values of the trait (except for the binary trait T2D), and β is the effect size estimated using untransformed values. The P values were obtained from the quantile-transformed-value-based analyses. All P values $<1 \times 10^{-6}$ are marked in bold and were all successfully replicated in the B99 cohort (Extended Data Table 3). Individuals with known T2D were removed from the analysis of quantitative traits. Highly significant associations were also observed when removing both individuals with known and screen-detected T2D from the analyses (data not shown). BMI, body mass index; HDL, high-density lipoprotein; HOMA-IR, homeostasis model assessment-estimated insulin resistance.

carriers with carriers of other genotypes. These analyses demonstrated that homozygous carriers of p.Arg684Ter in the IHIT had a 3.8 mmol l^{-1} higher 2-h plasma glucose level ($P_{\text{recessive model (rec)}} = 2.5 \times 10^{-35}$) (Table 1). Although the main effect was seen when comparing homozygous p.Arg684Ter carriers with all other individuals, even heterozygous carriers displayed a 0.43 mmol l^{-1} higher 2-h plasma glucose level than non-carriers ($P = 5.3 \times 10^{-5}$) (Fig. 3b). To further investigate the metabolic implications of p.Arg684Ter, we analysed additional metabolic traits (Table 1). Analyses of 220 individuals with T2D and 1,810 non-diabetic control individuals showed a strong association of p.Arg684Ter with increased risk of T2D ($P_{\text{additive model (add)}} = 2.1 \times 10^{-11}$). Similar to the 2-h plasma glucose levels, the data suggested a recessive inheritance for T2D (OR$_{\text{rec}}$, 10.3; $P_{\text{rec}} = 1.6 \times 10^{-24}$) (Fig. 3b). Interestingly, when using an alternative definition of T2D that is based on recent HbA$_{1C}$ criteria and does not include plasma glucose data, the association was modest ($P = 0.0084$). This finding is in line with the modest association of p.Arg684Ter with HbA$_{1C}$ as a quantitative trait (Table 1). We also found that p.Arg684Ter was associated with decreased peripheral insulin sensitivity, as estimated by the Gutt insulin sensitivity index (ISI)[10] (ISI$_{0,120}$: $\beta_{\text{add}} = -0.32$ s.d., $P_{\text{add}} = 1.4 \times 10^{-20}$; $\beta_{\text{rec}} = -1.0$ s.d., $P_{\text{rec}} = 1.6 \times 10^{-27}$) (Table 1). We replicated these findings in the B99 cohort and found consistent results (Extended Data Table 3). Finally, we found associations of p.Arg684Ter with lower fasting plasma glucose and serum insulin levels in the IHIT cohort but with substantially lower effect sizes than the glucose-stimulated effects ($\beta_{\text{add}} = -0.048$ mmol l^{-1}, $P_{\text{add}} = 0.00011$; $\beta_{\text{rec}} = -0.18$ mmol l^{-1}, $P_{\text{rec}} = 1.1 \times 10^{-6}$; and $\beta_{\text{add}} = -2.3$ pmol l^{-1}, $P_{\text{add}} = 0.00012$; $\beta_{\text{rec}} = -8.3$ pmol l^{-1}, $P_{\text{rec}} = 0.0014$, respectively) (Table 1). Thus, our findings indicate that the p.Arg684Ter *TBC1D4* variant confers increased risk of a subset of diabetes that features deterioration of postprandial glucose homeostasis. In this context, it is of interest that 2-h glucose levels appear to be a better predictor of cardiovascular disease than do fasting plasma glucose levels[11]. The p.Arg684Ter variant showed no significant association with HOMA-B, an estimate of basal β-cell function. Similarly, no convincing associations were detected with measures of adiposity, fasting lipid levels or other components of metabolic syndrome (Table 1).

The impact of p.Arg684Ter in its recessive form on 2-h plasma glucose (3.8 mmol l^{-1}) and T2D risk (OR, 10.3) are several times larger than any effects that have been reported in large-scale genome-wide association studies for these traits[2–4]. Furthermore, the p.Arg684Ter polymorphism has a high population impact, as 3.8% of Greenlanders are homozygous carriers of the risk allele. In the IHIT cohort, 15.5% of the patients with T2D are homozygous carriers of the risk allele, in contrast to 1.6% among glucose-tolerant individuals, indicating that p.Arg684Ter accounts for more than 10% of all cases of T2D in Greenland. Between 40 and 60 years of age, more than 60% of the homozygous carriers have T2D, and this increases to more than 80% above the age of 60. The effect of p.Arg864Ter thus mirrors a Mendelian-disease-like pattern of inheritance.

The p.Arg684Ter variant has a minor allele frequency (MAF) of 17% in the IHIT cohort, and we estimated it to have a MAF of 23% and 0% in the unobserved Inuit and European populations that are ancestral populations to the Greenlanders. In comparison, this variant was found in only 1 Japanese individual (NA18989) out of the 1,092 individuals sequenced in the 1000 Genomes Project[12], and it was not present in exome sequencing data from 2,000 Danish individuals[13], 448 Han Chinese individuals or ~6,500 European and African American individuals[14]. Thus, the variant is not unique to the Greenlandic population but is probably common only among Greenlanders and other related populations. This finding raises the question of whether the variant has been favoured by natural selection or whether it has increased in frequency as a result of genetic drift. A test for selection showed weak evidence for positive selection (Extended Data Fig. 5 and Supplementary Notes 1).

TBC1D4, also known as AS160, acts as a mediator of insulin-stimulated Akt-induced glucose uptake through Rab-mediated regulation of GLUT4 mobilization[15]. *Tbc1d4*-knockout mice have decreased basal plasma glucose levels and are resistant to insulin-stimulated glucose uptake in muscle and adipose tissue[16]. Furthermore, the overall GLUT4 levels in these mice are markedly lower than those of *Tbc1d4*-sufficient mice[16]. Two isoforms of the *TBC1D4* gene have been reported: one encodes a full-length protein, and the other encodes a short form lacking exons 11 and 12 (Fig. 3c). It is predicted that the p.Arg684Ter variant results in termination of *TBC1D4* transcription in exon 11; thus, this variant is expected to affect only the long isoform. In line with previous findings[17], expression analyses in humans showed that while the short isoform of *TBC1D4* is widely expressed, the long isoform is primarily expressed in skeletal muscle and not in other major tissues associated with glucose metabolism, such as adipose tissue, the liver or the pancreatic islets (Extended Data Fig. 6a, b). Thus, it is unlikely that p.Arg684Ter affects the latter tissues. We measured the expression levels in skeletal muscle tissue in groups of three individuals carrying zero, one or two copies of p.Arg684Ter. As predicted, the levels of long *TBC1D4* isoform mRNA and protein decreased with increasing number of p.Arg684Ter alleles (Fig. 3d, e). The short isoform was observed at very low levels in skeletal muscle regardless of the genotype (Extended Data Fig. 6c), indicating that this isoform is unlikely to contribute to the observed phenotype. Further analyses showed that GLUT4 protein levels in the skeletal muscle decreased with increasing number of p.Arg684Ter alleles (Extended Data Fig. 6d). Thus, the phenotype of global *Tbc1d4*-knockout mice—lower fasting glucose levels and markedly lower insulin-stimulated glucose uptake than *Tbc1d4*-sufficient mice[16]—is comparable to the phenotype observed in homozygous carriers of the p.Arg684Ter variant. Our data indicate that disruption of the full-length TBC1D4 protein in skeletal muscle results in severely decreased insulin-stimulated glucose uptake, leading to postprandial hyperglycaemia, impaired glucose tolerance and T2D.

The effect of *TBC1D4* p.Arg684Ter is in line with a reported familial case of postprandial hyperinsulinaemia caused by a *TBC1D4* p.Arg363Ter variant[18]. However, the reported p.Arg363Ter variant affects both TBC1D4 isoforms and consequently many tissues. Moreover, this variant has a large effect on insulin-stimulated glucose uptake in heterozygous carriers but not on fasting glucose levels. By contrast, the p.Arg684Ter variant discovered here has a large effect only in homozygous carriers and is restricted to the long isoform of TBC1D4, thereby affecting TBC1D4 signalling in skeletal muscle but not in β-cells, the liver or adipose tissue. Furthermore, the p.Arg684Ter variant affects fasting glucose levels. Finally, the high frequency of the nonsense variant in Greenlanders has allowed us to assess the physiological impact of this variant with high statistical confidence.

In summary, our study demonstrates the strength of conducting genetic association mapping outside the traditional setting of large homogeneous populations. We report a novel association of a common *TBC1D4* nonsense variant with T2D and elevated circulating glucose and insulin levels after an oral glucose load. The effect sizes of the variant markedly exceed previously reported associations of common genetic variants with metabolic traits. The variant leads to a prematurely terminated transcript of the long isoform of *TBC1D4*, which in homozygous carriers causes insulin resistance in skeletal muscle and confers a high risk of a subtype of T2D that is characterized by a deterioration of postprandial glucose homeostasis.

METHODS SUMMARY

Discovery association analyses were performed on the IHIT cohort data[6] and replicated with the B99 cohort data[9] (Extended Data Table 1). Participants underwent an oral glucose tolerance test, with plasma glucose and serum insulin levels measured at fasting and after 2 h. Diabetes was classified according to the World Health Organization. Samples were genotyped using the Cardio-Metabochip (Illumina)[7] with standard protocols. Samples with mis-specified gender, high rates of missing data and duplicates were removed using the software toolset PLINK. For association testing, we used a linear mixed model, implemented in the software GEMMA, to control for admixture and relatedness. Only participants without known diabetes were analysed (IHIT/B99: fasting plasma glucose, $n = 2{,}575/n = 1{,}064$; fasting serum insulin, $n = 2{,}575/n = 1{,}062$; 2-h plasma glucose, $n = 2{,}540/n = 845$; and 2-h serum insulin, $n = 2{,}540/n = 845$). All quantitative traits were quantile-transformed to a standard normal distribution; age and sex were included as covariates, and tests were performed using a likelihood ratio test. Effect sizes and their standard errors were estimated using a restricted maximum likelihood approach. Conditional analyses were performed by including the SNP that we conditioned on as an additional covariate. A meta-analysis was performed using the inverse-variance-based method. To estimate the OR for T2D under the recessive model, we used a linear mixed model without covariates, estimated α (the intercept) and β (the genotype effect) and set $\mathrm{OR} = \{(\alpha+\beta)/[1-(\alpha+\beta)]\}/[\alpha/(1-\alpha)]$. After the discovery studies, we selected nine trios that we inferred to have no European ancestry, enriching for carriers of the rs7330796 minor allele. We exome-sequenced the trios by using SureSelect capture (Agilent) followed by HiSeq2000 sequencing (Illumina), aligned the data using bwa software and called genotypes using SAMtools software. Variants that were in high linkage disequilibrium with rs7330796 were identified, genotyped in all individuals and tested for association. Admixture proportions were estimated using the software ADMIXTURE, assuming that there are two ancestral populations. Relatedness was estimated using the software RelateAdmix, which takes admixture into account. Methods for the biological studies are described in Methods.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 9 January; accepted 30 April 2014.

Published online 18 June 2014.

1. Jørgensen, M. E., Bjerregaard, P. & Borch-Johnsen, K. Diabetes and impaired glucose tolerance among the Inuit population of Greenland. *Diabetes Care* **25,** 1766–1771 (2002).
2. Scott, R. A. *et al.* Large-scale association analyses identify new loci influencing glycemic traits and provide insight into the underlying biological pathways. *Nature Genet.* **44,** 991–1005 (2012).
3. Morris, A. P. *et al.* Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nature Genet.* **44,** 981–990 (2012).
4. Manning, A. K. *et al.* A genome-wide approach accounting for body mass index identifies genetic variants influencing fasting glycemic traits and insulin resistance. *Nature Genet.* **44,** 659–669 (2012).
5. Rafnar, T. *et al.* Mutations in *BRIP1* confer high risk of ovarian cancer. *Nature Genet.* **43,** 1104–1107 (2011).
6. Jørgensen, M. E., Borch-Johnsen, K., Stolk, R. & Bjerregaard, P. Fat distribution and glucose intolerance among Greenland Inuit. *Diabetes Care* **36,** 2988–2994 (2013).
7. Voight, B. F. *et al.* The Metabochip, a custom genotyping array for genetic studies of metabolic, cardiovascular, and anthropometric traits. *PLoS Genet.* **8,** e1002793 (2012).
8. Marchini, J., Cardon, L. R., Phillips, M. S. & Donnelly, P. The effects of human population structure on large genetic association studies. *Nature Genet.* **36,** 512–517 (2004).
9. Bjerregaard, P. *et al.* Inuit health in Greenland: a population survey of life style and disease in Greenland and among Inuit living in Denmark. *Int. J. Circumpolar Health* **62** (suppl. 1), 3–79 (2003).
10. Gutt, M. *et al.* Validation of the insulin sensitivity index ($\mathrm{ISI}_{0,120}$): comparison with other measures. *Diabetes Res. Clin. Pract.* **47,** 177–184 (2000).
11. Cederberg, H. *et al.* Postchallenge glucose, A1C, and fasting glucose as predictors of type 2 diabetes and cardiovascular disease: a 10-year prospective cohort study. *Diabetes Care* **33,** 2077–2083 (2010).
12. The 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature* **467,** 1061–1073 (2010).
13. Lohmueller, K. E. *et al.* Whole-exome sequencing of 2,000 Danish individuals and the role of rare coding variants in type 2 diabetes. *Am. J. Hum. Genet.* **93,** 1072–1086 (2013).
14. NHLBI. *NHLBI GO Exome Sequencing Project (Exome Variant Server)* http://evs.gs.washington.edu/EVS/.
15. Sano, H. *et al.* Insulin-stimulated phosphorylation of a Rab GTPase-activating protein regulates GLUT4 translocation. *J. Biol. Chem.* **278,** 14599–14602 (2003).
16. Wang, H. Y. *et al.* AS160 deficiency causes whole-body insulin resistance via composite effects in multiple tissues. *Biochem. J.* **449,** 479–489 (2013).
17. Baus, D. *et al.* Identification of a novel AS160 splice variant that regulates GLUT4 translocation and glucose-uptake in rat muscle cells. *Cell. Signal.* **20,** 2237–2246 (2008).
18. Dash, S. *et al.* A truncation mutation in *TBC1D4* in a family with acanthosis nigricans and postprandial hyperinsulinemia. *Proc. Natl Acad. Sci. USA* **106,** 9350–9355 (2009).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank I. Kleist, as well as all colleagues at the Physician's Clinic, Nuuk, Greenland. We also thank J. F. Wojtaszewski for comments and X. Zhou for discussions about GEMMA. The Novo Nordisk Foundation Center for Basic Metabolic Research is an independent research centre at the University of Copenhagen and is partly funded by an unrestricted donation from the Novo Nordisk Foundation. This project was also funded by the Danish Council for Independent Research (Medical Sciences), the Steno Diabetes Center and the Villum Foundation. The IHIT study was supported by Karen Elise Jensen's Foundation, NunaFonden, the Medical Research Council of Denmark, the Medical Research Council of Greenland and the Commission for Scientific Research in Greenland. None of the funding agencies had any role in the study design or in the collection or interpretation of the data.

Author Contributions T.H. and A.A. conceived and headed the project. I.M., R.N. and A.A. designed the statistical set-up, and T.H., N.G., A.P.G. and O.P. designed the experimental set-up for the DNA extraction, genotyping and sequencing. A.L. provided the Danish samples. M.E.J., P.B. and K.B.-J. provided the Greenlandic samples, collected and defined the phenotypes and provided context for these samples. I.M. and A.A. performed the admixture, relatedness and linkage disequilibrium analyses. I.M. carried out the statistical part of the association analysis and N.G. carried out the medical part, with input from A.A., T.H., O.P. and R.N. The Chinese samples were analysed by X.W., G.S., X.J., J.A.-A. and J.W. N.G. analysed the Danish samples. X.W., G.S., X.J., J.A.-A. and J.W. performed the library constructions and sequencing. T.S.K. performed the mapping and genotyping for the sequencing data. M.F. and R.N. performed the selection analysis. N.T.K. collected muscle biopsies, and J.T.T., T.S.N., M.A.A. and J.R.Z. experimentally analysed and interpreted the *TBC1D4* and *GLUT4* expression data. N.G., I.M., A.A. and T.H. wrote most of the manuscript, with input from R.N., O.P., M.E.J. and P.B. All authors approved the final version of the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to T.H. (torben.hansen@sund.ku.dk) or A.A. (albrecht@binf.ku.dk).

doi:10.1038/nature13408

Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA

Emilia Huerta-Sánchez[1,2,3]*, Xin Jin[1,4]*, Asan[1,5,6]*, Zhuoma Bianba[7]*, Benjamin M. Peter[2], Nicolas Vinckenbosch[2], Yu Liang[1,5,6], Xin Yi[1,5,6], Mingze He[1,8], Mehmet Somel[9], Peixiang Ni[1], Bo Wang[1], Xiaohua Ou[1], Huasang[1], Jiangbai Luosang[1], Zha Xi Ping Cuo[10], Kui Li[11], Guoyi Gao[12], Ye Yin[1], Wei Wang[1], Xiuqing Zhang[1,13,14], Xun Xu[1], Huanming Yang[1,15,16], Yingrui Li[1], Jian Wang[1,16], Jun Wang[1,15,17,18,19] & Rasmus Nielsen[1,2,20,21]

As modern humans migrated out of Africa, they encountered many new environmental conditions, including greater temperature extremes, different pathogens and higher altitudes. These diverse environments are likely to have acted as agents of natural selection and to have led to local adaptations. One of the most celebrated examples in humans is the adaptation of Tibetans to the hypoxic environment of the high-altitude Tibetan plateau[1–3]. A hypoxia pathway gene, *EPAS1*, was previously identified as having the most extreme signature of positive selection in Tibetans[4–10], and was shown to be associated with differences in haemoglobin concentration at high altitude. Re-sequencing the region around *EPAS1* in 40 Tibetan and 40 Han individuals, we find that this gene has a highly unusual haplotype structure that can only be convincingly explained by introgression of DNA from Denisovan or Denisovan-related individuals into humans. Scanning a larger set of worldwide populations, we find that the selected haplotype is only found in Denisovans and in Tibetans, and at very low frequency among Han Chinese. Furthermore, the length of the haplotype, and the fact that it is not found in any other populations, makes it unlikely that the haplotype sharing between Tibetans and Denisovans was caused by incomplete ancestral lineage sorting rather than introgression. Our findings illustrate that admixture with other hominin species has provided genetic variation that helped humans to adapt to new environments.

The Tibetan plateau (at greater than 4,000 m) is inhospitable to human settlement because of low atmospheric oxygen pressure (~40% lower than at sea level), cold climate and limited resources (for example, sparse vegetation). Despite these extreme conditions, Tibetans have successfully settled in the plateau, in part due to adaptations that confer lower infant mortality and higher fertility than acclimated women of low-altitude origin. The latter tend to have difficulty bearing children at high altitude, and their offspring typically have low birth weights compared to offspring from women of high-altitude ancestry[1,2]. One well-documented pregnancy-related complication due to high altitude is the higher incidence of preeclampsia[2,11] (hypertension during pregnancy). In addition, the physiological response to low oxygen differs between Tibetans and individuals of low-altitude origin. For most individuals, acclimatization to low oxygen involves an increase in blood haemoglobin levels. However, in Tibetans, the increase in haemoglobin levels is limited[3], presumably because high haemoglobin concentrations are associated with increased blood viscosity and increased risk of cardiac events, thus resulting in a net reduction in fitness[12,13].

Recently, the genetic basis underlying adaptation to high altitude in Tibetans was elucidated[4–10] using exome and single nucleotide polymorphism (SNP) array data. Several genes seem to be involved in the response but most studies identified *EPAS1*, a transcription factor induced under hypoxic conditions, as the gene with the strongest signal of Tibetan specific selection[4–10]. Furthermore, SNP variants in *EPAS1* showed significant associations with haemoglobin levels in the expected direction in several of these studies; individuals carrying the derived allele have lower haemoglobin levels than individuals homozygous for the ancestral allele. Here, we re-sequence the complete *EPAS1* gene in 40 Tibetan and 40 Han individuals at more than 200× coverage to further characterize this impressive example of human adaptation. Remarkably, we find the source of adaptation was likely to be due to the introduction of genetic variants from archaic Denisovan-like individuals (individuals closely related to the Denisovan individual from the Altai Mountains[14]) into the ancestral Tibetan gene pool.

After applying standard next-generation sequencing filters (see Methods), we call a total of 477 SNPs in a region of approximately 129 kilobases (kb) in the combined Han and Tibetan samples (Supplementary Tables 1 and 2). We compute the fixation index (F_{ST}; see Methods) between Han and Tibetans, and confirm that it is highly elevated in the *EPAS1* region as expected under strong local selection (Extended Data Fig. 1). Indeed, by comparison to 26 populations from the Human Genome Diversity Panel[15,16] (Fig. 1) it is clear that the variants in this region are far more differentiated than one would expect given the average genome-wide differentiation between Han and Tibetans (F_{ST} ~0.02, ref. 4). The only other genes with comparably large frequency differences between any closely related populations are the previously identified loci associated with lighter skin pigmentation in Europeans, *SLCA45A2* and *HERC2* (refs 17–20), although in these examples the populations compared (for example, Hazara and French, Brahui and Russians) are more genetically differentiated than Han and Tibetans. In populations as closely related as Han and Tibetans, we find no examples of SNPs with as much differentiation as seen in *EPAS1*, illustrating the uniqueness of its selection signal.

F_{ST} is particularly elevated in a 32.7-kb region containing the 32 most differentiated SNPs (green box in Extended Data Fig. 1 and Supplementary Table 3), which is the best candidate region for the advantageous mutation(s). We therefore focus the subsequent analyses primarily on this region. Phasing the data (see Methods) to identify Han and Tibetan haplotypes in this region (Fig. 2), we find that Tibetans carry a high-frequency haplotype pattern that is strikingly different from both their

[1]BGI-Shenzhen, Shenzhen 518083, China. [2]Department of Integrative Biology, University of California, Berkeley, California 94720 USA. [3]School of Natural Sciences, University of California, Merced, California 95343 USA. [4]School of Bioscience and Bioengineering, South China University of Technology, Guangzhou 510006, China. [5]Binhai Genomics Institute, BGI-Tianjin, Tianjin 300308, China. [6]Tianjin Translational Genomics Center, BGI-Tianjin, Tianjin 300308, China. [7]The People's Hospital of Lhasa, Lhasa 850000, China. [8]Bioinformatics and Computational Biology Program, Iowa State University, Ames, Iowa 50011, USA. [9]Department of Biological Sciences, Middle East Technical University, 06800 Ankara, Turkey. [10]The Second People's Hospital of Tibet Autonomous Region, Lhasa 850000, China. [11]The People's Hospital of the Tibet Autonomous Region, Lhasa 850000, China. [12]The hospital of XiShuangBanNa Dai Nationalities, Autonomous Jinghong, 666100 Yunnan, China. [13]The Guangdong Enterprise Key Laboratory of Human Disease Genomics, BGI-Shenzhen, 518083 Shenzhen, China. [14]Shenzhen Key Laboratory of Transomics Biotechnologies, BGI-Shenzhen, 518083 Shenzhen, China. [15]Princess Al Jawhara Center of Excellence in the Research of Hereditary Disorders, King Abdulaziz University, Jeddah 21589, Saudi Arabia. [16]James D. Watson Institute of Genome Science, 310008 Hangzhou, China. [17]Department of Biology, University of Copenhagen, Ole MaaløesVej 5, 2200 Copenhagen, Denmark. [18]Macau University of Science and Technology, AvenidaWai long, Taipa, Macau 999078, China. [19]Department of Medicine, University of Hong Kong 999077, Hong Kong. [20]Department of Statistics, University of California, Berkeley, California 94720, USA. [21]Department of Biology, University of Copenhagen, 2200 Copenhagen, Denmark.
*These authors contributed equally to this work.

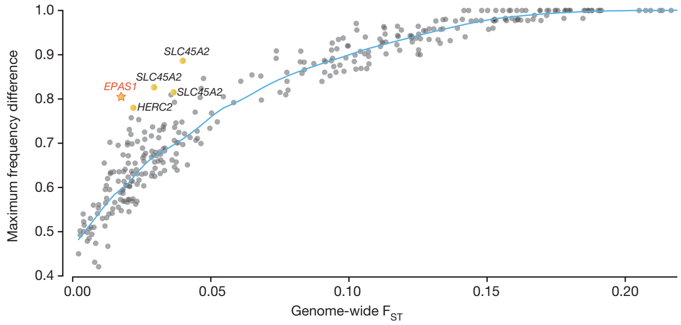


Figure 1 | Genome-wide F_{ST} versus maximal allele frequency difference. The relationship between genome-wide F_{ST} (x axis) computed for each pair of the 26 populations and maximal allele frequency difference (y axis), first explored in ref. 19. Maximal allele frequency difference is defined as the largest frequency difference observed for any SNP between a population pair. The 26 populations are from the Human Genome Diversity Panel (HGDP). The labels highlight genes that harbour SNPs previously identified as having strong local adaptation. The grey points represent the observed relationship between population differentiation (F_{ST}) and maximal allele frequency difference; the more differentiated populations tend to have mutations with larger frequency differences. The star symbol and the yellow symbols represent outliers; these are populations that are not highly differentiated but where we find some mutations that have higher frequency differences than expected (light blue line).

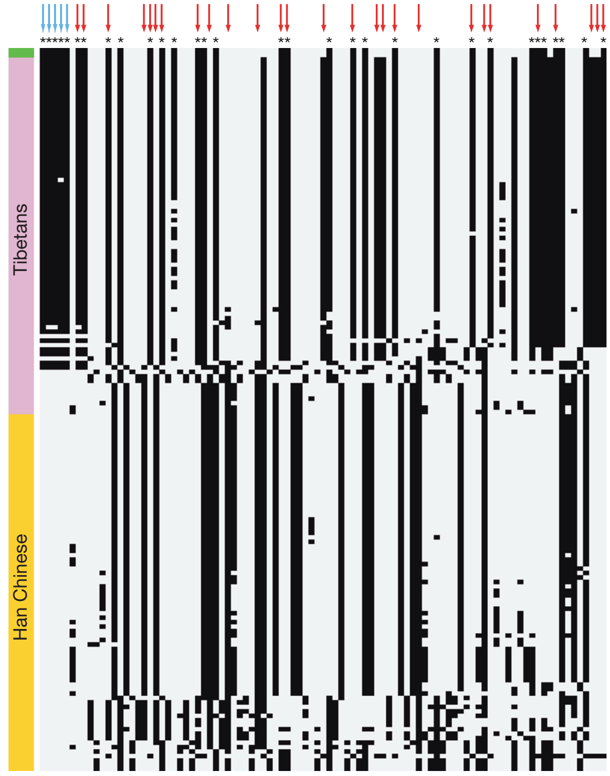


Figure 2 | Haplotype pattern in a region defined by SNPs that are at high frequency in Tibetans and at low frequency in Han Chinese. Each column is a polymorphic genomic location (95 in total), each row is a phased haplotype (80 Han and 80 Tibetan haplotypes), and the coloured column on the left denotes the population identity of the individuals. Haplotypes of the Denisovan individual are shown in the top two rows (green). The black cells represent the presence of the derived allele and the grey space represents the presence of the ancestral allele (see Methods). The first and last columns correspond to the first and last positions in Supplementary Table 3, respectively. The red and blue arrows indicate the 32 sites in Supplementary Table 3. The blue arrows represent a five-SNP haplotype block defined by the first five SNPs in the 32.7-kb region. Asterisks indicate sites at which Tibetans share a derived allele with the Denisovan individual.

minority haplotypes and the common haplotype observed in Han Chinese For example, the region harbours a highly differentiated 5-SNP haplotype motif (AGGAA) within a 2.5-kb window that is only seen in Tibetan samples and in none of the Han samples (the first five SNPs in Supplementary Table 3, and blue arrows in Fig. 2). The pattern of genetic variation within Tibetans appears even more unusual because none of the variants in the five-SNP motif is present in any of the minority haplotypes of Tibetans. Even when subject to a selective sweep, we would not generally expect a single haplotype to contain so many unique mutations not found on other haplotypes.

We investigate whether a model of selection on either a *de novo* mutation (SDN) or selection on standing variation (SSV) could possibly lead to so many fixed differences between haplotype classes in such a short region within a single population. To do so, we simulate a 32.7-kb region under these models assuming different strengths of selection and conditioning on the current allele frequency in the sample (see Methods). We find that the observed number of fixed differences between the haplotype classes is significantly higher than what is expected by simulations under any of the models explored (Extended Data Fig. 2). Thus the degree of differentiation between haplotypes is significantly larger than expected from mutation, genetic drift and directional selection alone. In other words, it is unlikely ($P < 0.02$ under either a SSV scenario or under a SDN scenario) that the high degree of haplotype differentiation could be caused by a single beneficial mutation landing by chance on a background of rare SNPs, which are then brought to high frequency by selection. The remaining explanations are the presence of strong epistasis between many mutations, or that a divergent population introduced the haplotype into Tibetans by gene flow or through ancestral lineage sorting.

We search for potential donor populations in two different data sets: the 1000 Genomes Project[21] and whole genome data from ref. 14. We originally defined the *EPAS1* 32.7-kb region boundaries by the level of observed differentiation between the Tibetans and Han only (Supplementary Table 3, Extended Data Fig. 1 and Fig. 2) as described in the previous section. In that region, the most common haplotype in Tibetans is tagged by the distinctive five-SNP motif (AGGAA; the first five SNPs in Fig. 2), not found in any of our 40 Han samples. We first focus on this five-SNP motif and determine whether it is unique to Tibetans or if it is found in other populations.

Intriguingly, when we examine the 1000 Genomes Project data set, we discover that the Tibetan five-SNP motif (AGGAA) is not present in any of these populations, except for a single CHS (Southern Han Chinese) and a single CHB (Beijing Han Chinese) individual. Extended Data Fig. 3 contains the frequencies of all the haplotypes present in the fourteen 1000 Genomes populations[21] at these five SNP positions. Furthermore, when we examine the data set from ref. 14 containing both modern (Papuan, San, Yoruba, Mandeka, Mbuti, French, Sardinian, Han Dai, Dinka, Karitiana, and Utah residents of northern and western European ancestry (CEU)) and archaic (high-coverage Denisovan and low-coverage Croatian Neanderthal) human genomes[14], we discover that the five-SNP motif is completely absent in all of their modern human population samples (Supplementary Table 4). Therefore, apart from one CHS and one CHB individual, none of the other extant human populations sampled to date carry this five-SNP haplotype. Notably, the Denisovan haplotype at these five sites (AGGAA) exactly matches the five-SNP Tibetan motif (Supplementary Table 4 and Extended Data Fig. 3).

We observe the same pattern when focusing on the entire 32.7-kb region and not just the five-SNP motif. Twenty SNPs in this region have unusually high frequency differences of at least 0.65 between Tibetans and all the other populations from the 1000 Genomes Project (Extended Data Fig. 4). However, in Tibetans, 15 out of these 20 SNPs are identical to the Denisovan haplotype generating an overall pattern of high haplotype similarity between the selected Tibetan haplotype and the Denisovan haplotype (Supplementary Tables 5–7). Interestingly, five of these SNPs in the region are private SNPs shared between Tibetans and the Denisovan, but not shared with any other population worldwide, except for two SNPs at low frequency in Han Chinese (Extended Data Fig. 4 and Supplementary Table 7).

If we consider all SNPs (not just the most differentiated) in the 32.7-kb region annotated in humans, to build a haplotype network[22] using the 40 most common haplotypes, we observe a clear pattern in which the Tibetan haplotype is much closer to the Denisovan haplotype than any modern human haplotype (Fig. 3 and Extended Fig. 5a; see Extended Data Fig. 6a, b for haplotype networks constructed using other criteria). Furthermore, we find that the Tibetan haplotype is slightly more divergent from other modern human populations than the Denisovan haplotype is, a pattern expected under introgression (see Methods and Extended Data Fig. 5b). Raw sequence divergence for all sites and all haplotypes shows a similar pattern (Extended Data Fig. 7). Moreover, the divergence between the common Tibetan haplotype and Han haplotypes is larger than expected for comparisons among modern humans, but well within the distribution expected from human–Denisovan comparisons (Extended Data Fig. 8). Notably, sequence divergence between the Tibetans' most common haplotype and Denisovan is significantly lower ($P = 0.0028$) than expected from human–Denisovan comparisons (Extended Data Fig. 8). We also find that the number of pairwise differences between the common Tibetan haplotype and the Denisovan haplotype ($n = 12$) is compatible with the levels one would expect from mutation accumulation since the introgression event (see Methods for Extended Data Fig. 8). Finally, if we compute D (ref. 14) and S^* (refs 23, 24), two statistics that have been designed to detect archaic introgression into modern humans, we obtain significant values (D-statistic $P < 0.001$, and S^* $P \leq 0.035$) for the 32.7-kb region using multiple null models of no gene flow (see Methods, Supplementary Tables 8–10, and Extended Data Figs 9 and 10a).

Thus, we conclude that the haplotype associated with altitude adaptation in Tibetans is likely to be a product of introgression from Denisovan or Denisovan-related populations. The only other possible explanation is ancestral lineage sorting. However, this explanation is very unlikely as it cannot explain the significant D and S^* values and because it would require a long haplotype to be maintained without recombination since the time of divergence between Denisovans and humans (estimated to be at least 200,000 years (ref. 14)). The chance of maintaining a 32.7-kb fragment in both lineages throughout 200,000 years is conservatively estimated at $P = 0.0012$ assuming a constant recombination of 2.3×10^{-8} per base pair (bp) per generation (see Methods). Furthermore, the haplotype would have to have been independently lost in all African and non-African populations, except for Tibetans and Han Chinese.

We have re-sequenced the *EPAS1* region and found that Tibetans harbour a highly differentiated haplotype that is only found at very low frequency in the Han population among all the 1000 Genomes populations, and is otherwise only observed in a previously sequenced Denisovan individual[14]. As the haplotype is observed in a single individual in both CHS and CHB samples, it suggests that it was introduced into humans before the separation of Han and Tibetan populations, but subject to selection in Tibetans after the Tibetan plateau was colonized. Alternatively, recent admixture from Tibetans to Hans may have introduced the haplotype to nearby Han populations outside Tibet. The CHS and CHB individuals carrying the five-SNP Tibetan–Denisovan haplotype (Extended Data Fig. 3) show no evidence of being recent migrants from Tibet (see Methods and Extended Data Fig. 10b), suggesting that if the haplotype was carried from Tibet to China by migrants, this migration did not occur within the last few generations.

Previous studies examining the genetic contributions of Denisovans to modern humans[14,25] suggest that Melanesians have a much larger Denisovan component than either Han or Mongolians, even though the latter populations are geographically much closer to the Altai mountains[14,25]. Interestingly, the putatively beneficial Denisovan *EPAS1* haplotype is not observed in modern-day Melanesians or in the high-coverage Altai Neanderthal[26] (Supplementary Table 4). Evidence has been found for

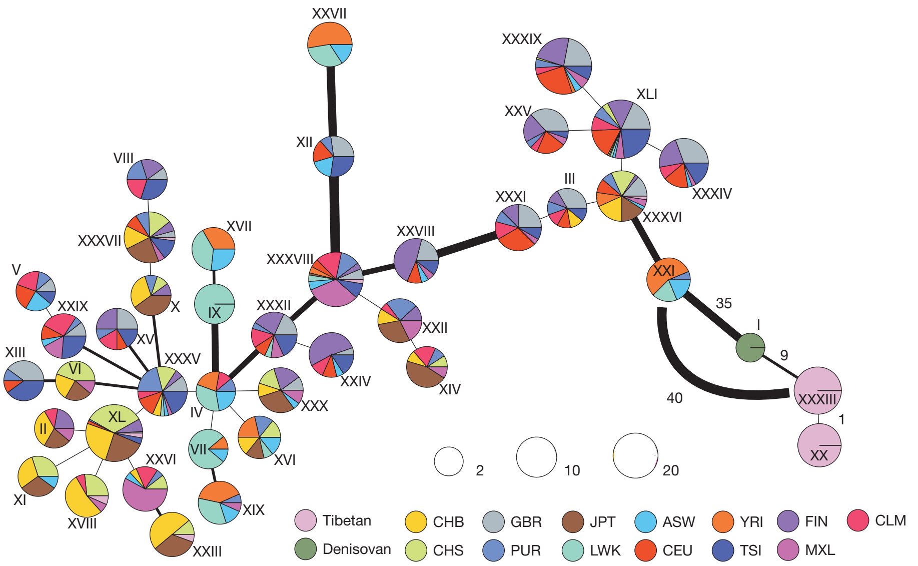


Figure 3 | A haplotype network based on the number of pairwise differences between the 40 most common haplotypes. The haplotypes were defined from all the SNPs present in the combined 1000 Genomes and Tibetan samples: 515 SNPs in total within the 32.7-kb *EPAS1* region. The Denisovan haplotypes were added to the set of the common haplotypes. The R software package pegas[23] was used to generate the figure, using pairwise differences as distances. Each pie chart represents one unique haplotype, labelled with Roman numerals, and the radius of the pie chart is proportional to the log$_2$(number of chromosomes with that haplotype) plus a minimum size so that it is easier to see the Denisovan haplotype. The sections in the pie provide the breakdown of the haplotype representation amongst populations. The width of the edges is proportional to the number of pairwise differences between the joined haplotypes; the thinnest edge represents a difference of one mutation. The legend shows all the possible haplotypes among these populations. The numbers (1, 9, 35 and 40) next to an edge (the line connecting two haplotypes) in the bottom right are the number of pairwise differences between the corresponding haplotypes. We added an edge afterwards between the Tibetan haplotype XXXIII and its closest non-Denisovan haplotype (XXI) to indicate its divergence from the other modern human groups. Extended Data Fig. 5a contains all the pairwise differences between the haplotypes presented in this figure. ASW, African Americans from the south western United States; CEU, Utah residents with northern and western European ancestry; GBR, British; FIN, Finnish; JPT, Japanese; LWK, Luhya; CHS, southern Han Chinese; CHB, Han Chinese from Beijing; MXL, Mexican; PUR, Puerto Rican; CLM, Colombian; TSI, Toscani; YRI, Yoruban. Where there is only one line within a pie chart, this indicates that only one population contains the haplotype.

Denisovan admixture throughout southeast Asia (as well as in Melanesians) based on a global analysis of SNP array data from 1,600 individuals from a diverse set of populations[27], and this finding has been recently confirmed by ref. 26. Therefore, it appears that sufficient archaic admixture into populations near the Tibetan region occurred to explain the presence of this Denisovan haplotype outside Melanesia. Furthermore, the haplotype may have been maintained in some human populations, including Tibetans and their ancestors, through the action of natural selection.

Recently, a few studies have supported the idea of adaptive introgression from archaic humans to modern humans as having a role in the evolution of immunity-related genes (HLA (ref. 28) and *STAT2* (ref. 29)) and in the evolution of skin pigmentation genes (*BNC2* (refs 23, 30)). Our findings imply that one of the most clear-cut examples of human adaptation is likely to be due to a similar mechanism of gene flow from archaic hominins into modern humans. With our increased understanding that human evolution has involved a substantial amount of gene flow from various archaic species, we are now also starting to understand that adaptation to local environments may have been facilitated by gene flow from other hominins that may already have been adapted to those environments.

METHODS SUMMARY

DNA samples included in this work were extracted from peripheral blood of 41 unrelated Tibetan individuals living at more than 4,300 m above sea level within the Himalayan Plateau, with informed consent. Tibetan identity was based on self-reported family ancestry. The individuals were from two villages of Dingri (4,300 m altitude) and Naqu (4,600 m altitude). These individuals are a subset of the 50 individuals exome-sequenced analysed in ref. 4. Samples of 40 Han Chinese (CHB) are from the 1000 Genomes Project. A combined strategy of long-range PCR and next-generation sequencing was used to decipher the whole *EPAS1* gene and its ±30-kb flanking region. We designed 38 pairs of long-range PCR primers to amplify the region in 41 Tibetan and 40 Han individuals. PCR products from all individuals were fragmented and indexed, then sequenced to higher than 260-fold depth for each individual with the Illumina Hiseq2000 sequencer. The reads were aligned to the UCSC human reference genome (hg18) using the SOAPaligner. Genotypes of each individual at every genomic location of the *EPAS1* gene and the flanking region were called by SOAPsnp. To make comparisons with the 40 Han easier, we only used 40 Tibetan samples for this study.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 17 December 2013; accepted 28 April 2014.

Published online 2 July; corrected online 13 August 2014 (see full-text HTML version for details).

1. Moore, L. G., Young, D., McCullough, R. E., Droma, T. & Zamudio, S. Tibetan protection from intrauterine growth restriction (IUGR) and reproductive loss at high altitude. *Am. J. Hum. Biol.* **13,** 635–644 (2001).
2. Niermeyer, S. *et al.* Child health and living at high altitude. *Arch. Dis. Child.* **94,** 806–811 (2009).
3. Wu, T. *et al.* Hemoglobin levels in Quinghai-Tibet: different effects of gender for Tibetans vs. Han. *J. Appl. Physiol.* **98,** 598–604 (2005).
4. Yi, X. *et al.* Sequencing of 50 human exomes reveals adaptation to high altitude. *Science* **329,** 75–78 (2010).
5. Bigham, A. *et al.* Identifying signature of natural selection in Tibetan and Andean populations using dense genome scan data. *PLoS Genet.* **6,** e1001116 (2010).
6. Simonson, T. S. *et al.* Genetic evidence for high-altitude adaptation in Tibet. *Science* **329,** 72–75 (2010).
7. Beall, C. M. *et al.* Natural selection on EPAS1 (HIF2a) associated with low hemoglobin concentration in Tibetan highlanders. *Proc. Natl Acad. Sci. USA* **107,** 11459–11464 (2010).
8. Peng, Y. *et al.* Genetic variations in Tibetan populations and high-altitude adaptation at the Himalayas. *Mol. Biol. Evol.* **28,** 1075–1081 (2011).
9. Xu, S. *et al.* A genome-wide search for signals of high-altitude adaptation in Tibetans. *Mol. Biol. Evol.* **28,** 1003–1011 (2011).
10. Wang, B. *et al.* On the origin of Tibetans and their genetic basis in adapting high-altitude environments. *PLoS ONE* **6,** e17002 (2011).
11. Moore, L. G. *et al.* Maternal adaptation to high-altitude pregnancy: an experiment of nature—a review. *Placenta* **25,** S60–S71 (2004).
12. Vargas, E. & Spielvogel, H. Chronic mountain sickness, optimal hemoglobin, and heart disease. *High Alt. Med. Biol.* **7,** 138–149 (2006).
13. Yip, R. Significance of an abnormally low or high hemoglobin concentration during pregnancy: special consideration of iron nutrition1'2'3. *Am. J. Clin. Nutr.* **72,** 272S–279S (2000).
14. Meyer, M. *et al.* A high-coverage genome sequence from an archaic Denisovan individual. *Science* **338,** 222–226 (2012).
15. Li, J. Z. *et al.* Worldwide human relationships inferred from genome-wide patterns of variation. *Science* **319,** 1100–1104 (2008).
16. Rosenberg, N. A. Standardized subsets of the HGDP-CEPH Human Genome Diversity Cell Line Panel, accounting for atypical and duplicated samples and pairs of close relatives. *Ann. Hum. Genet.* **70,** 841–847 (2006).
17. Soejima, M. & Koda, Y. Population differences of two coding SNPs. in pigmentation-related genes SLC24A5 and SLC45A2. *Int. J. Legal Med.* **121,** 36–39 (2007).
18. Sulem, P. *et al.* Genetic determinants of hair, eye and skin pigmentation in Europeans. *Nature Genet.* **39,** 1443–1452 (2007).
19. Coop, G. *et al.* The role of geography in human adaptation. *PLoS Genet.* **5,** e1000500 (2009).
20. Pickrell, J. K. *et al.* Signals of recent positive selection in a worldwide sample of human populations. *Genome Res.* **19,** 826–837 (2009).
21. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491,** 56–65 (2012).
22. Paradis, E. Pegas: an R package for population genetics with an integrated–modular approach. *Bioinformatics* **26,** 419–420 (2010).
23. Vernot, B. & Akey, J. Resurrecting Surviving neandertal lineages from modern human genomes. *Science* (2014).
24. Plagnol, V. & Wall, J. D. Possible ancestral structure in human populations. *PLoS Genet.* **2,** e105 (2006).
25. Reich, D. *et al.* Genetic history of an archaic hominin group from Denisova cave in Siberia. *Nature* **468,** 1053–1060 (2010).
26. Prüfer, K. *et al.* The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature* **505,** 43–49 (2014).
27. Skoglund, P. & Jakobsson, M. Archaic human ancestry in East Asia. *Proc. Natl Acad. Sci. USA* **108,** 18301–18306 (2011).
28. Abi-Rached, L. *et al.* The shaping of modern human immune systems by multiregional admixture with archaic humans. *Science* **334,** 89–94 (2011).
29. Mendez, F. L., Watkins, J. C. & Hammer, M. F. A haplotype at STAT2 introgressed from Neanderthals and serves as a candidate of positive selection in Papua New Guinea. *Am. J. Hum. Genet.* **91,** 265–274 (2012).
30. Sankararaman, S. *et al.* The genomic landscape of Neanderthal ancestry in present-day humans. *Nature* (2014).

Supplementary Information is available in the online version of the paper.

Acknowledgements This research was funded by the State Key Development Program for Basic Research of China, 973 Program (2011CB809203, 2012CB518201, 2011CB809201, 2011CB809202), China National GeneBank-Shenzhen and Shenzhen Key Laboratory of Transomics Biotechnologies (no. CXB201108250096A). This work was also supported by research grants from the US NIH; R01HG003229 to R.N. and R01HG003229-08S2 to E.H.S. We thank F. Jay, M. Liang and F. Casey for useful discussions.

Author Contributions R.N., Ji.W. and Ju.W. supervised the project. X.J., A., Z.B., Y.L., X.Y., M.H., P.N., B.W., X.O., H., J.L., Z.X.P.C., K.L., G.G., Y.Y., W.W., X.Z., X.X., H.Y., Y.L., Ji.W. and Ju.W. collected and generated the data, and performed the preliminary bioinformatic analyses to call SNPs and indels from the raw data. E.H.-S. and N.V. filtered the data and B.M.P. phased the data. E.H.-S. performed the majority of the population genetic analysis with some contributions from B.M.P. and M.S. E.H.-S. and R.N. wrote the manuscript with critical input from all the authors.

Author Information Sequence data have been deposited in the Sequence Read Archive under accession number SRP041218. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to Ju.W. (wangj@genomics.cn) or Ji.W. (wangjian@genomics.cn) or R.N. (rasmus_nielsen@berkeley.edu).

doi:10.1038/nature13619

Replication stress is a potent driver of functional decline in ageing haematopoietic stem cells

Johanna Flach[1,2], Sietske T. Bakker[1], Mary Mohrin[1], Pauline C. Conroy[3], Eric M. Pietras[1], Damien Reynaud[1], Silvia Alvarez[4], Morgan E. Diolaiti[5], Fernando Ugarte[6], E. Camilla Forsberg[6], Michelle M. Le Beau[7], Bradley A. Stohr[5], Juan Méndez[4], Ciaran G. Morrison[3] & Emmanuelle Passegué[1]

Haematopoietic stem cells (HSCs) self-renew for life, thereby making them one of the few blood cells that truly age[1,2]. Paradoxically, although HSCs numerically expand with age, their functional activity declines over time, resulting in degraded blood production and impaired engraftment following transplantation[2]. While many drivers of HSC ageing have been proposed[2–5], the reason why HSC function degrades with age remains unknown. Here we show that cycling old HSCs in mice have heightened levels of replication stress associated with cell cycle defects and chromosome gaps or breaks, which are due to decreased expression of mini-chromosome maintenance (MCM) helicase components and altered dynamics of DNA replication forks. Nonetheless, old HSCs survive replication unless confronted with a strong replication challenge, such as transplantation. Moreover, once old HSCs re-establish quiescence, residual replication stress on ribosomal DNA (rDNA) genes leads to the formation of nucleolar-associated γH2AX signals, which persist owing to ineffective H2AX dephosphorylation by mislocalized PP4c phosphatase rather than ongoing DNA damage. Persistent nucleolar γH2AX also acts as a histone modification marking the transcriptional silencing of rDNA genes and decreased ribosome biogenesis in quiescent old HSCs. Our results identify replication stress as a potent driver of functional decline in old HSCs, and highlight the MCM DNA helicase as a potential molecular target for rejuvenation therapies.

Both human and mouse HSCs accumulate γH2AX signals with age[6,7]. This is taken as direct evidence of DNA damage occurring in old HSCs, since phosphorylation of histone H2AX by ATM or ATR upon sensing of DNA breaks is one of the first steps in the canonical DNA damage response (DDR)[8]. The idea that DNA damage is a driver of HSC ageing is also supported by the age-related functional impairment observed in HSCs isolated from mice deficient in DNA repair pathway components[6,9]. Accumulation of DNA damage in old HSCs is an attractive hypothesis to explain the propensity of the ageing blood system to acquire mutations[10], especially since quiescent HSCs are particularly vulnerable to genomic instability after DNA damage, owing to their preferential use of the error-prone non-homologous end joining (NHEJ) repair pathway[11]. However, it remains to be established what causes γH2AX accumulation with age, and how it contributes to the functional decline of old HSCs.

To address these questions, we isolated HSCs as $\mathrm{Lin}^-/\mathrm{cKit}^+/\mathrm{Sca1}^+/\mathrm{Flk2}^-/\mathrm{CD48}^-/\mathrm{CD150}^+$ cells from the bone marrow of young (6–12 weeks) and old (22–30 months) wild-type C57BL/6 mice (Extended Data Fig. 1a). We confirmed the functional impairment of old HSCs compared with young HSCs, with the expected reduced engraftment, loss of lymphoid

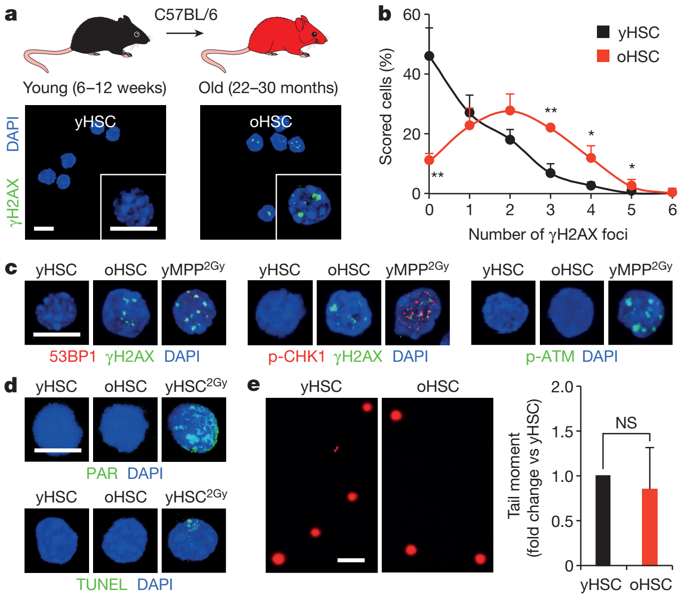


Figure 1 | Accumulation of γH2AX foci without detectable DNA damage in old HSCs. **a**, **b**, Representative images (**a**) and quantification (**b**) of γH2AX foci in young and old HSCs (yHSC and oHSC, respectively). **c**, **d**, Representative images of DNA damage markers in young and old HSCs: **c**, 53BP1 or p-CHK1 recruitment in γH2AX foci and p-ATR activation; **d**, PAR detection and TUNEL staining. 2 Gy irradiated cells ($\mathrm{yMPP}^{2\mathrm{Gy}}$ and $\mathrm{yHSC}^{2\mathrm{Gy}}$) are included as positive controls. **e**, Representative images of young and old HSCs analysed by alkaline comet assay and quantification of mean tail moment ($n = 4$). Results are expressed as fold change compared with young HSCs (set to 1). Scale bars, 10 μm (**a**, **c**, **d**); 90 μm (**e**). Data are means ± standard deviation (s.d.). $*P \le 0.05$, $**P \le 0.01$. NS, not significant.

[1] The Eli and Edythe Broad Center for Regenerative Medicine and Stem Cell Research, Department of Medicine, Hem/Onc Division, University of California San Francisco, San Francisco, California 94143, USA. [2] Institute of Experimental Cancer Research, Comprehensive Cancer Center, 89081 Ulm, Germany. [3] Center for Chromosome Biology, School of Natural Sciences, National University of Ireland Galway, Galway, Ireland. [4] Spanish National Cancer Research Centre (CNIO), E-28049 Madrid, Spain. [5] Department of Pathology, University of California San Francisco, San Francisco, California 94143, USA. [6] Institute for the Biology of Stem Cells, Department of Biomolecular Engineering, University of California Santa Cruz, Santa Cruz, California 95064, USA. [7] Section of Hematology/Oncology and the Comprehensive Cancer Center, University of Chicago, Chicago, Illinois 60637, USA.

potential and early onset of bone marrow failure or myeloid malignancies following transplantation (Extended Data Fig. 1b)[2,5]. We also confirmed that old HSCs contain more γH2AX signals than young HSCs (Fig. 1a, b and Extended Data Fig. 2a)[6]. However, we found no evidence of associated co-localization of DNA damage proteins by microscopy, or DNA fragmentation by poly-ADP-ribose (PAR) and TdT-mediated dUTP nick end labelling (TUNEL) staining (Fig. 1c, d and Extended Data Fig. 2b, c). We also performed alkaline comet assays to directly measure the number of DNA breaks and, although both populations showed some very damaged outliers, no statistical difference in mean tail moment was observed between young and old HSCs (Fig. 1e and Extended Data Fig. 2d, e). Importantly, we tested the effect of 0.5 Gy of ionizing radiation on young HSCs, since this dose was estimated to be equivalent to the level of γH2AX signals present in old HSCs[6], and observed increased tail moment by comet assay and 53BP1/γH2AX co-localization, hence validating the sensitivity of our assays (Extended Data Fig. 2f, g). We also found that age-associated γH2AX signals were considerably less intense than ionizing-radiation-induced γH2AX foci (Extended Data Fig. 3a), which probably reflects differences in the spread and density of phosphorylated H2AX in each case. Collectively, these results indicate that old HSCs display γH2AX signals without DDR activation or detectable levels of DNA breaks.

To determine whether old HSCs remain competent for DDR, we exposed young and old HSCs to 2 Gy of ionizing radiation and followed their kinetics of DNA repair by microscopy (Fig. 2a and Extended Data Fig. 3b). In both populations, we observed increased 53BP1-containing γH2AX foci by 2 h after ionizing radiation, followed by their progressive disappearance over time. Although old HSCs showed slower kinetics, both populations had essentially cleared all ionizing-radiation-induced γH2AX foci by 24 h after irradiation (Fig. 2b). In addition, both young and old HSCs expressed equivalent levels of homologous recombination and NHEJ DNA repair genes by quantitative polymerase chain reaction with reverse transcription (qRT–PCR) analyses (Fig. 2c). Altogether, these results demonstrate that old HSCs can activate the DDR and clear ionizing-radiation-induced γH2AX foci as effectively as young HSCs, and suggest that accumulation of γH2AX in old HSCs could be independent of the sensing of DNA breaks. In fact, ATR can also be activated upon sensing interference with DNA replication forks[8]. Strikingly, we observed increased staining for the single-stranded DNA-binding proteins RPA and ATRIP in old HSCs (Fig. 2d, e and Extended Data Fig. 3c), which suggests that age-associated γH2AX signals could originate from replication stress[12].

Replication stress is intrinsically linked to cell proliferation, and previous studies have reported a spectrum of findings ranging from increased, decreased, to unchanged proliferation in old HSCs[13]. In our hands, cell cycle analyses revealed a variable frequency of G0/G1 cells in old HSCs (Fig. 3a and Extended Data Fig. 4a). qRT–PCR analyses of cell cycle genes also indicated enhanced expression of *Cdkn1a* (p21) and decreased expression of a range of cyclins in old HSCs, which suggest engagement of cell cycle restriction checkpoints (Extended Data Fig. 4b). Moreover, tracking of single-cell division kinetics uncovered a consistent ~4 h delay in

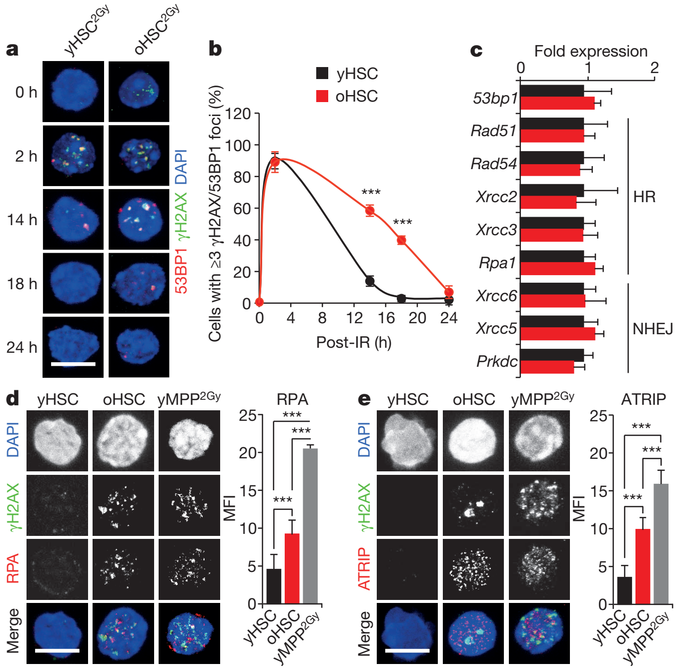


Figure 2 | Efficient DNA repair but persistence of replication stress remnants in old HSCs. **a**, **b**, Representative images (**a**) and quantification (**b**) of DNA repair kinetics in 2 Gy irradiated young and old HSCs (yHSC^{2Gy} and oHSC^{2Gy}, respectively; $n = 3$). IR, ionizing radiation. **c**, qRT–PCR analyses of HR and NHEJ gene expression in young and old HSCs ($n = 4$). Results are expressed as fold change compared with young HSCs (set to 1). **d**, **e**, Representative images and mean fluorescence intensity (MFI) quantification of RPA (**d**) and ATRIP (**e**) staining in young and old HSCs. 2 Gy irradiated cells are included as positive control. Scale bars, 10 μm. Data are means ± s.d. ***$P \leq 0.001$.

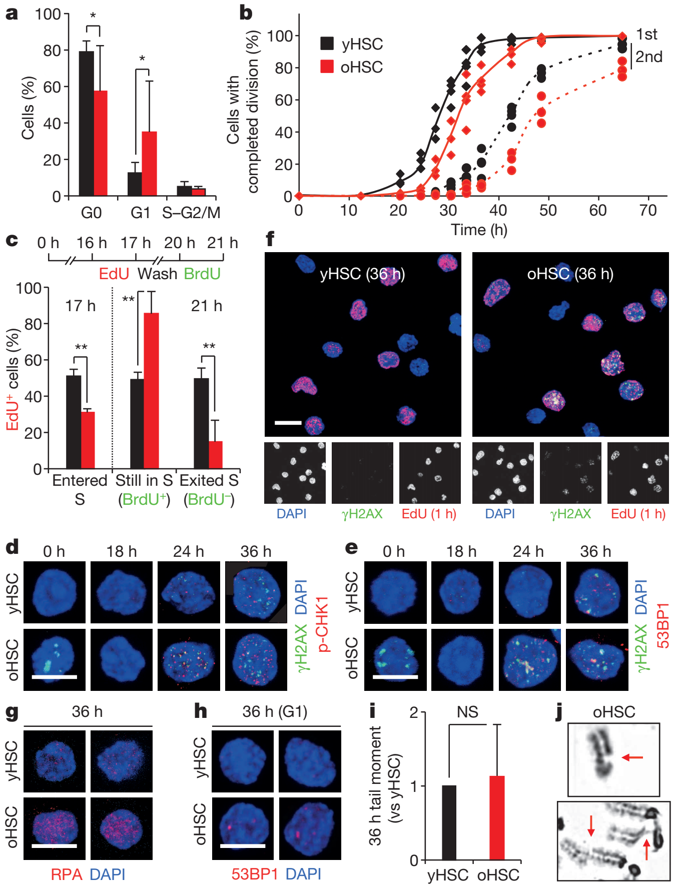


Figure 3 | Replication stress in cycling old HSCs. **a**, Cell cycle distribution of young and old HSCs (yHSC and oHSC, respectively; $n = 8$). **b**, Single cell tracking to measure the kinetics of the first and second cell division in cultured young and old HSCs ($n = 3$). **c**, EdU and EdU/BrdU labelling of cycling young and old HSCs ($n = 4$). **d**, **e**, Representative images of γH2AX/p-CHK1 (**d**) and γH2AX/53BP1 (**e**) foci in cycling young and old HSCs. **f**, Representative images of γH2AX/EdU staining in 36 h cycling young and old HSCs. **g**, **h**, Representative images of RPA staining (**g**) and persistent G1-phase 53BP1 bodies (**h**) in 36 h cycling young and old HSCs. **i**, Quantification of mean tail moment in 36 h cycling young and old HSCs by alkaline comet assay ($n = 4$). Results are expressed as fold change compared with yHSCs (set to 1). **j**, Representative reverse image of DAPI-stained metaphase cell from 5-day-expanded old HSCs showing chromatid gaps (arrows). Scale bars, 10 μm. Data are means ± s.d. *$P \leq 0.05$, **$P \leq 0.01$, ***$P \leq 0.001$. NS, not significant.

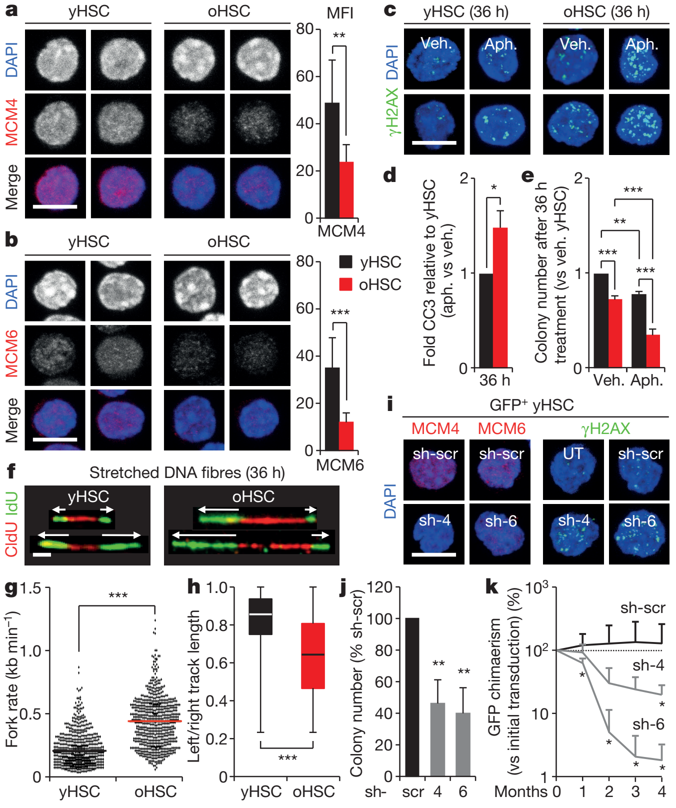


Figure 4 | Defective replication due to reduced MCM expression in old HSCs. **a**, **b**, Representative images and quantification (MFI) of MCM4 (**a**) and MCM6 (**b**) protein levels in young and old HSCs (yHSC and oHSC, respectively). **c**–**e**, Effect of low-dose aphidicolin (Aph.; 50 ng ml^{-1}) on cultured young and old HSCs ($n = 3$): **c**, representative images of γH2AX foci; **d**, cleaved caspase 3 (CC3) levels; and **e**, colony counts in methylcellulose after 36 h treatment. Results are normalized for vehicle-treated cells (Veh.) and expressed as fold change compared with young HSCs (set to 1). **f**–**h**, Analyses of CldU/IdU-labelled DNA replication tracks in 36 h cycling young and old HSCs ($n = 3$): **f**, representative images (arrows indicate fork progression); **g**, individual fork velocities with means (bars); and **h**, box plot quantification of fork symmetry ratio. **i**–**k**, Lentiviral-mediated knockdown of *Mcm4* and *Mcm6* in young HSCs ($n = 3$). Transduced green fluorescent protein (GFP)$^{+}$ HSCs were re-isolated 48 h post-infection for *in vitro* analyses, or used without re-isolation 12 h post-infection for transplantation (5 mice per condition): **i**, representative images of MCM4 and MCM6 protein levels and γH2AX foci; **j**, colony counts in methylcellulose (results are expressed as fold change compared to scrambled shRNA (sh-scr)-infected HSCs, set to 100%); and **k**, reconstitution ability upon transplantation (results are percentage of GFP chimaerism normalized to the initial transduction efficiency per construct). UT, untransfected. Scale bars, 10 μm (**a**–**c**, **i**); 2.5 μm (**f**). Data are means ± s.d. $*P \leq 0.05$, $**P \leq 0.01$, $***P \leq 0.001$.

the onset of the first division in old HSCs, which was even increased during the second division (Fig. 3b). We directly confirmed that old HSCs had both a delayed entry into S phase and an extended S phase using single 5-ethynyl-2′-deoxyuridine (EdU) and double EdU/5-bromodeoxyuridine (BrdU) incorporation experiments (Fig. 3c and Extended Data Fig. 4c). Collectively, these results demonstrate impaired progression through S phase in cycling old HSCs.

DNA replication is often accompanied by γH2AX foci formation at stalled and/or collapsed replication forks, and activation of the DDR to allow normal DNA synthesis[14]. Strikingly, we found that cycling old HSCs displayed significantly more phosphorylated (p)-CHK1 and 53BP1-containing γH2AX foci than young HSCs, and directly showed γH2AX accumulation in replicating old HSCs using EdU/γH2AX co-staining of both *in vitro* and *in vivo* cycling cells (Fig. 3d–f and Extended Data Fig. 4c, d). We also confirmed elevated RPA staining in 36 h cycling old HSCs (Fig. 3g), and persistent 53BP1 bodies in ~60% of old HSCs that had re-entered G1 phase at this time point compared with ~20% of young HSCs (Fig. 3h and Extended Data Fig. 4e)[15]. Moreover, we observed a trend towards elevated amounts of fragmented DNA detected by alkaline comet assays in cycling old HSCs either cultured *in vitro* (not significant) or isolated after *in vivo* mobilization treatment (Fig. 3i and Extended Data Fig. 4f, g). Consistently, 5-day-expanded cultures showed increased numbers of chromosomal gaps and breaks in the progeny of cycling old HSCs (Fig. 3j and Supplementary Table 1). In contrast, karyotyping analyses revealed no evidence of chromosomal deletions and/or translocations (Supplementary Table 1), which usually occur as a consequence of NHEJ-mediated repair of DNA breaks[11]. Collectively, these results demonstrate heightened levels of replication stress in cycling old HSCs associated with extended S phase and acquisition of chromosomal gaps/breaks. In rare cases, we also observed exacerbated features of replication stress in old HSCs, including senescence with increased senescence-associated β-galactosidase (SA-β-Gal) staining and *Cdkn2a* (p16) expression, and fragile telomeres with multiple telomeric signals (Extended Data Fig. 5a, b).

To understand at the molecular level why old HSCs have replication stress, we performed microarray gene expression analyses. We compared both HSCs and granulocyte/macrophage progenitors (GMPs) and subtracted for genes that were differentially expressed between young and old GMPs. This allowed us to identify 913 significantly differentially expressed genes that were specific to old HSCs and segregated into four main clusters (Supplementary Table 2). Among those, we observed a selective downregulation of all MCM genes (*Mcm2–7*), which encode the six subunits of the MCM DNA helicase (Extended Data Fig. 5c, d)[16]. Using qRT–PCR, we confirmed unchanged levels of *Mcm* genes in old GMPs, and significantly decreased expression of at least *Mcm4* and *Mcm6* in both quiescent and cycling old HSCs (Extended Data Figs 5e and 6a, b). Moreover, we directly demonstrated a ~50% decrease in MCM4 and MCM6 protein levels in quiescent old HSCs (Fig. 4a, b). Interestingly, re-analyses of published data sets also showed decreased *Mcm* expression in several old HSC samples (Extended Data Fig. 6c). The MCM proteins form a heterohexameric complex that is part of the pre-replication complex assembled at origins of replication during late M/early G1 phases[16]. At the G1-to-S-phase transition, MCM proteins associate with CDC45 and the go-ichi-ni-san (GINS) complex to form an active helicase that unwinds the DNA at replication forks[17]. In contrast to MCM proteins, the expression of other pre-replication complex or DNA helicase components was not altered in old HSCs (Extended Data Fig. 6a). Collectively, these results uncovered a specific deficit in MCM proteins in old HSCs.

One characteristic of cells with decreased MCM levels is hypersensitivity to replication stressors[18]. To test the sensitivity of young and old HSCs, we used a low dose of the DNA polymerase inhibitor aphidicolin. While 36 h *in vitro* aphidicolin treatment only had a modest effect on young HSCs, old HSCs displayed a massive accumulation of γH2AX foci, enhanced apoptosis and severely impaired colony-forming activity upon re-plating in methylcellulose (Fig. 4c–e). However, after transplantation, 36 h aphidicolin-treated young HSCs showed strikingly impaired reconstitution ability, with early onset of bone marrow failure and death, while both treated and untreated old HSCs displayed equally poor transplantability with reduced lymphoid output (Extended Data Fig. 7a). Moreover, the number of engrafted 36 h aphidicolin-treated young HSCs was significantly reduced to levels similar to engrafted old HSCs, either treated or untreated (Extended Data Fig. 7b). These findings demonstrate that induced replication stress severely damages the functionality of young HSCs in a way that resembles age-associated effects, and that transplantation is the ultimate replication challenge for old HSCs. We also confirmed that the differential killing of old HSCs was specific for replication stressors as opposed to non-specific cytotoxic agents (Extended Data Fig. 7c). Collectively, these results demonstrate that replication stress can degrade HSC function, even in young HSCs with a full complement of MCM proteins, and that old HSCs with reduced MCM levels are more susceptible to the killing effect of replication challenges both *in vitro* and *in vivo*.

Although MCM proteins are normally present in excess, downregulation of just one component is sufficient to sensitize cells to replication stress by reducing their capacity to activate dormant origins in response to collapsed replication forks[18]. To directly assess replication at the single-molecule level, we analysed stretched DNA fibres after 5-chloro-2′-deoxyuridine (CldU)/5-iodo-2′-deoxyuridine (IdU) labelling of 36 h replicating young and old HSCs (Fig. 4f and Extended Data Fig. 7d). Strikingly, we observed a significant increase in both replication fork velocity and numbers of asymmetric replication forks in old HSCs (Fig. 4h, g). These altered dynamics are consistent with the replication stress features reported in cells with impaired MCM activity[18,19], and probably reflect a reduced number of licensed replication origins leading to an extended S phase in old HSCs. We then used lentiviral vectors containing either *Mcm4* or *Mcm6* short hairpin RNA (shRNA) to determine the effect of decreased MCM levels on young HSC function. We confirmed ≥50% knockdown both at the messenger RNA and protein levels, associated with accumulation of γH2AX foci, lower colony-forming ability and reduced expansion rates in culture (Fig. 4i, j and Extended Data Figs 7e, f, 8a), hence demonstrating impaired replication in transduced young HSCs. Moreover, transplantation experiments showed decreased reconstitution ability from transduced HSCs (Fig. 4k), which directly confirms that reducing MCM levels strongly impairs young HSC functionality. Collectively, these results reinforce previous data linking decreased MCM levels to impaired stem- and progenitor-cell proliferation[20], and identify the deficit in MCM DNA helicase components as the likely cause of the replication stress features of old HSCs.

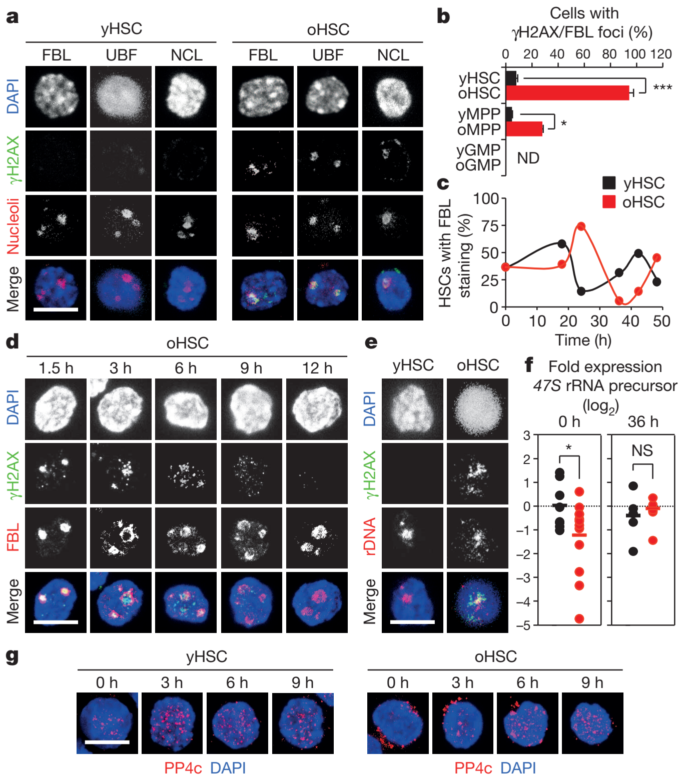


Figure 5 | Persistent nucleolar γH2AX foci in quiescent old HSCs. **a**, Representative images of γH2AX and nucleolar marker co-localization in young and old HSCs (yHSC and oHSC, respectively): fibrillarin (FBL); upstream binding factor (UBF); and nucleolin (NCL). **b**, Quantification of γH2AX/FBL foci in young and old cells. ND, not detectable. **c**, Representative kinetics of nucleolar dissociation/reformation in cultured young and old HSCs ($n = 3$). **d**, Representative images of FBL/γH2AX staining in cultured old HSCs. **e**, Representative images of immuno-FISH for γH2AX and rDNA in young and old HSCs. **f**, qRT–PCR analyses of *47S* rRNA precursor transcript expression in quiescent ($n = 12$) and cycling ($n = 8$) young and old HSCs. Results are expressed as $\log_2$ fold change compared with young HSCs (set to 0). Bars indicate average expression levels. **g**, Representative images of PP4c staining in cultured young and old HSCs. Scale bars, 10 μm. Data are means ± s.d. $*P \leq 0.05$, $***P \leq 0.001$. NS, not significant.

While replication stress provides an explanation for the high levels of γH2AX in cycling old HSCs, it does not account for γH2AX accumulation in quiescent old HSCs. To address this issue, we asked whether γH2AX could mark particular subnuclear structures[21]. Whereas no co-localization was found with centromeric or telomeric regions, we observed an almost complete co-localization of γH2AX signals with nucleolar markers in quiescent old HSCs (Fig. 5a and Extended Data Fig. 8b, c)[22]. Although nucleolar γH2AX signals were almost never found in young cells, they were occasionally observed in old multipotent progenitors (MPPs), but not in old GMPs or granulocytes (Fig. 5b and Extended Data Fig. 8d, e). One to five nucleoli can usually be observed per mouse cell, which result from the cell-cycle-dependent assembly of nucleolar organizer regions (NORs) present on four different chromosome pairs in the C57BL/6 mouse background[22]. As expected, replicating HSCs showed a cyclical dissociation/reformation of nucleolar structures, albeit with slower kinetics in old HSCs (Fig. 5c and Extended Data Fig. 8f). Remarkably, γH2AX signals quickly vanished from the nucleolus in cycling old HSCs even before nucleolar dissociation, and never re-appeared *in vitro* even after nucleolar reformation (Fig. 5d and Extended Data Fig. 8f). Nucleolar γH2AX signals were also never observed in cycling old HSCs re-isolated 2 weeks after transplantation, but were readily detectable in old HSCs re-isolated 7 months after transplantation, which had, by then, re-entered quiescence (Extended Data Fig. 9a). The nucleolus is primarily the site of ribosome biogenesis, where multiple repeats of rDNA genes present on each NOR are transcribed and then spliced to produce the 18S, 5.8S and 28S rRNA subunits[22]. We confirmed the presence of γH2AX on rDNA genes in quiescent old HSCs using an immunofluorescence *in situ* hybridization (FISH) approach (Fig. 5e and Extended Data Fig. 8g). We also found significantly reduced expression of the *47S* rRNA precursor transcripts by qRT–PCR in quiescent old HSCs (Fig. 5f), and confirmed decreased ribosome biogenesis in these cells using Bioanalyzer track analyses (Extended Data Fig. 9b). Moreover, we observed restoration of *47S* rRNA precursor transcript expression to levels found in young HSCs in cycling old HSCs that had lost nucleolar γH2AX (Fig. 5f). Taken together, these results indicate that nucleolar γH2AX signals are an exclusive feature of quiescent old HSCs, which correlate with decreased ribosome biogenesis and could mark the transcriptional silencing of rDNA genes. In contrast, none of the classic histone methylation marks associated with active or repressed transcription displayed specific nucleolar enrichment in old HSCs (Extended Data Fig. 9c).

rDNA genes are the most abundant and highly transcribed genes in eukaryotes, which contain many replication origins and are known to challenge the replication machinery[23]. It is therefore likely that replicating old HSCs accumulate γH2AX on rDNA genes, and we propose that their aggregation during nucleolar reformation causes the appearance of nucleolar-associated γH2AX signals in quiescent old HSCs. Importantly, single-nucleotide polymorphism (SNP) analyses of amplified genomic DNA did not reveal significant differences in rDNA sequences between young and old HSCs (data not shown), which suggests that replication stress has little to no mutagenic consequence for rDNA gene integrity. In addition, we found that PP4c, one of the best-characterized γH2AX phosphatases[24], was strikingly mislocalized in quiescent old HSCs. While nuclear PP4c was observed in both quiescent and cycling young HSCs, PP4c was found almost exclusively in the cytoplasm of quiescent old HSCs and only became nuclear when old HSCs re-entered the cell cycle (Fig. 5g and Extended Data Fig. 9d). Nuclear re-localization of PP4c also occurred within the same time window (3–9 h) as disappearance of γH2AX from the nucleolus in cycling old HSCs. Thus, it is conceivable that the long-term persistence of nucleolar γH2AX in quiescent old HSCs results from ineffective H2AX dephosphorylation due to mislocalized PP4c rather than ongoing DNA damage. Although we observed the presence of unrepaired, RPA-coated stretches of single-stranded DNA in quiescent old HSCs, they do not appear to trigger the DDR in these cells, in contrast to what has been described in other contexts[25]. They might also be the source of the DNA breaks detected by alkaline comet assays in a recent study of quiescent old HSCs[26], but again without evidence of an activated DDR. A failure to dephosphorylate H2AX could therefore explain why quiescent old HSCs show persistent γH2AX signals without DDR activation.

Our results demonstrate that replication stress is a potent driver of functional decline in old HSCs, and identify a deficit in MCM helicase components as the molecular mechanism for the impaired replication of old HSCs (Extended Data Fig. 10). It will now be important to understand why expression of *Mcm* genes decreases with age in HSCs, and whether this could be reversed through direct changes in old HSCs or rejuvenation of the ageing bone marrow niche[27]. Our results also suggest a non-canonical function for γH2AX as an epigenetic histone modification that marks the silencing of the transcription machinery. This could be a normal mechanism to block transcription in genomic regions undergoing DNA repair, and further studies will address the broad relevance of this novel finding. It will also be interesting to determine whether decreased rDNA gene transcription in quiescent old HSCs plays a part in bone marrow failure syndromes and other age-related blood defects linked to defective ribosome function[28]. In this context, it will be important to understand why PP4c is mislocalized in quiescent old HSCs, and whether this can be reverted for therapeutic purposes.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 3 July 2013; accepted 27 June 2014.

Published online 30 July 2014.

1. López-Otín, C., Blasco, M. A., Partridge, L., Serrano, M. & Kroemer, G. The hallmarks of aging. *Cell* **153,** 1194–1217 (2013).
2. Rossi, D. J., Jamieson, C. H. & Weissman, I. L. Stems cells and the pathways to aging and cancer. *Cell* **132,** 681–696 (2008).
3. Rossi, D. J. *et al.* Cell intrinsic alterations underlie hematopoietic stem cell aging. *Proc. Natl Acad. Sci. USA* **102,** 9194–9199 (2005).
4. Chambers, S. M. *et al.* Aging hematopoietic stem cells decline in function and exhibit epigenetic dysregulation. *PLoS Biol.* **5,** e201 (2007).
5. Geiger, H., de Haan, G. & Florian, M. C. The ageing haematopoietic stem cell compartment. *Nature Rev. Immunol.* **13,** 376–389 (2013).
6. Rossi, D. J. *et al.* Deficiencies in DNA damage repair limit the function of haematopoietic stem cells with age. *Nature* **447,** 725–729 (2007).
7. Rübe, C. E. *et al.* Accumulation of DNA damage in hematopoietic stem and progenitor cells during human aging. *PLoS ONE* **6,** e17487 (2011).
8. Ciccia, A. & Elledge, S. J. The DNA damage response: making it safe to play with knives. *Mol. Cell* **40,** 179–204 (2010).
9. Nijnik, A. *et al.* DNA repair is limiting for haematopoietic stem cells during ageing. *Nature* **447,** 686–690 (2007).
10. Welch, J. S. *et al.* The origin and evolution of mutations in acute myeloid leukemia. *Cell* **150,** 264–278 (2012).
11. Mohrin, M. *et al.* Hematopoietic stem cell quiescence promotes error prone DNA repair and mutagenesis. *Cell Stem Cell* **7,** 174–185 (2010).
12. Burhans, W. C. & Weinberger, M. DNA replication stress, genome instability and aging. *Nucleic Acids Res.* **35,** 7545–7556 (2007).
13. Pietras, E. M., Warr, M. R. & Passegué, E. Cell cycle regulation in hematopoietic stem cells. *J. Cell Biol.* **195,** 709–720 (2011).
14. Branzei, D. & Foiani, M. Maintaining genome stability at the replication fork. *Nature Rev. Mol. Cell Biol.* **11,** 208–219 (2010).
15. Lukas, C. *et al.* 53BP1 nuclear bodies form around DNA lesions generated by mitotic transmission of chromosomes under replication stress. *Nature Cell Biol.* **13,** 243–253 (2011).
16. Méndez, J. & Stillman, B. Chromatin association of human origin recognition complex, Cdc6 and minichromosome maintenance proteins during the cell cycle: assembly of prereplication complexes in late mitosis. *Mol. Cell. Biol.* **20,** 8602–8612 (2000).
17. Aparicio, T., Guillou, E., Coloma, J., Montoya, G. & Méndez, J. The human GINS complex associates with Cdc45 and MCM and is essential for DNA replication. *Nucleic Acids Res.* **37,** 2087–2095 (2009).
18. Ibarra, A., Schwob, E. & Méndez, J. Excess MCM proteins protect human cells from replicative stress by licensing backup origins of replication. *Proc. Natl Acad. Sci. USA* **105,** 8956–8961 (2008).
19. Zhong, Y. *et al.* The level of origin firing inversely affects the rate of replication fork progression. *J. Cell Biol.* **201,** 373–383 (2013).
20. Pruitt, S. C., Bailey, K. J. & Freeland, A. Reduced Mcm2 expression results in severe stem/progenitor cell deficiency and cancer. *Stem Cells* **25,** 3121–3132 (2007).
21. Lanctôt, C., Cheutin, T., Cremer, M., Cavalli, G. & Cremer, T. Dynamic genome architecture in the nuclear space: regulation of gene expression in three dimensions. *Nature Rev. Genet.* **8,** 104–115 (2007).
22. Boisvert, F. M., van Koningsbruggen, S., Navascués, J. & Lamond, A. I. The multifunctional nucleolus. *Nature Rev. Mol. Cell Biol.* **8,** 574–585 (2007).
23. Durkin, S. G. & Glover, T. W. Chromosome fragile sites. *Annu. Rev. Genet.* **41,** 169–192 (2007).
24. Nakada, S., Chen, G. I., Gingas, A.-C. & Durocher, D. PP4 is a γH2AX phosphatase required for recovery from the DNA damage checkpoint. *EMBO Rep.* **9,** 1019–1026 (2008).
25. Fumagalli, M. *et al.* Telomeric DNA damage is irreparable and causes persistent DNA-damage-response activation. *Nature Cell Biol.* **14,** 355–365 (2012).
26. Beerman, I., Seita, J., Inlay, M. A., Weissman, I. L. & Rossi, D. J. Quiescent hematopoietic stem cells accumulate DNA damage during aging that is repaired upon entry into cell cycle. *Cell Stem Cell* **15,** 37–50 (2014).
27. Rando, T. A. & Chang, H. Y. Aging, rejuvenation, and epigenetic reprogramming: resetting the aging clock. *Cell* **148,** 46–57 (2012).
28. Narla, A., Hurst, S. N. & Ebert, B. L. Ribosome defects in disorders of erythropoiesis. *Int. J. Hematol.* **93,** 144–149 (2011).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank A. Brunet and S. Villeda for providing some old C57BL/6 mice, B. McStay for advice on nucleolar analyses, C. Klijn for assistance with microarray analyses, S. Katzman for the SNP analyses, E. Davis for help with cytogenetic studies, I. Grummt for the rDNA plasmid, M. Kissner and M. Lee for management of our Flow Cytometry Core Facility, and all members of the Passegué laboratory for critical insights and suggestions. S.T.B. and M.M. were supported by a California Institute for Regenerative Medicine (CIRM) training grant and E.M.P. by National Institutes of Health (NIH) F32 HL106989. This work was supported by Science Foundation Ireland PI award 10/IN.1/B2972 to C.G.M. and a CIRM New Faculty Award RN2-00934 and NIH R01 HL092471 to E.P.

Author Contributions J.F. performed all of the experiments with help from S.T.B. for the comet assays and microarray data analyses, E.M.P. for the Ki67/DAPI staining and microarray analyses, and D.R. for the transplantation experiments. M.M. and P.C.C. initiated these studies. S.A. and J.M. performed the DNA replication track analyses and helped with the MCM experiments, M.E.D. and B.A.S. performed the telomere analyses, F.U. and E.C.F. performed the SNP analyses, and M.M.L.B. performed the cytogenetic breakage studies and spectral karyotyping analyses. J.F., C.G.M. and E.P. designed the experiments and interpreted the results. J.F. and E.P. wrote the manuscript.

Author Information Data have been deposited in the Gene Expression Omnibus under accession number GSE48893. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to E.P. (PassegueE@stemcell.ucsf.edu).

doi:10.1038/nature13410

Historical contingency and its biophysical basis in glucocorticoid receptor evolution

Michael J. Harms[1,2] & Joseph W. Thornton[2,3]

Understanding how chance historical events shape evolutionary processes is a central goal of evolutionary biology[1–7]. Direct insights into the extent and causes of evolutionary contingency have been limited to experimental systems[7–9], because it is difficult to know what happened in the deep past and to characterize other paths that evolution could have followed. Here we combine ancestral protein reconstruction, directed evolution and biophysical analysis to explore alternative 'might-have-been' trajectories during the ancient evolution of a novel protein function. We previously found that the evolution of cortisol specificity in the ancestral glucocorticoid receptor (GR) was contingent on permissive substitutions, which had no apparent effect on receptor function but were necessary for GR to tolerate the large-effect mutations that caused the shift in specificity[6]. Here we show that alternative mutations that could have permitted the historical function-switching substitutions are extremely rare in the ensemble of genotypes accessible to the ancestral GR. In a library of thousands of variants of the ancestral protein, we recovered historical permissive substitutions but no alternative permissive genotypes. Using biophysical analysis, we found that permissive mutations must satisfy at least three physical requirements—they must stabilize specific local elements of the protein structure, maintain the correct energetic balance between functional conformations, and be compatible with the ancestral and derived structures—thus revealing why permissive mutations are rare. These findings demonstrate that GR evolution depended strongly on improbable, non-deterministic events, and this contingency arose from intrinsic biophysical properties of the protein.

Historians and evolutionary biologists have long wrestled with the idea that historical outcomes may hinge on chance events. How differently would the world have turned out if the Persian cavalry had been present at the Battle of Marathon or if the KT asteroid had missed the Earth? In biology, evolutionary trajectories driven solely by the deterministic force of natural selection will always produce the optimal accessible form, irrespective of chance events[3,10]. In contrast, when non-deterministic processes such as drift play a strong part, the outcome depends on whatever chance events occur during evolution; if history could be set in motion again from some past starting point, very different results would probably unfold.

Recent studies show that the evolution of some protein functions was contingent on prior 'permissive' mutations, which are functionally neutral in isolation but must be present for the function-altering mutations to be tolerated[6,7,9,11–15]. Permissive mutations cannot be fixed by selection for the derived function and must therefore accumulate stochastically with respect to it. It remains unknown, however, how many permissive mutations could have enabled these evolutionary transitions and therefore whether the dependence on non-deterministic events is strong or weak. If the suite of potential permissive mutations is large, then many different evolutionary paths could enable the function-switching mutations, and the outcome of protein evolution would be only weakly contingent on its specific history. Conversely, if only a few mutations have the potential to permit the realized outcome, the probability that one of these would occur by chance would be very small, and the particular form and function achieved by the evolving protein would be strongly contingent on a low-probability event.

Understanding evolutionary contingency requires measuring the number of potentially permissive mutations and characterizing the factors that determine that number. Because history happened only once, this knowledge has been inaccessible for natural biological systems that evolved in the deep past. We addressed this issue by reconstructing ancestral proteins and subjecting them to directed evolution, a protein engineering strategy to efficiently characterize regions of protein sequence space with respect to some function of interest[16,17]. We then employed biophysical analyses to explore the mechanistic factors that determined the number of permissive genotypes.

We previously characterized an evolutionary transition in the GR ligand-binding domain (LBD) of bony vertebrates and found that it was contingent on permissive mutations[6]. The LBD serves as an allosterically regulated transcriptional activator: hormone binding causes the 'activation-function helix' (AF-H) to pack against the body of the protein, creating a new surface to which coactivator proteins bind, and increasing the transcription of nearby target genes[18,19]. Using ancestral protein reconstruction, we previously found that the cortisol-specific GR evolved from a promiscuous ancient receptor (AncGR1) because of seven historical substitutions that are conserved in all extant GRs (Fig. 1a, b)[6]. Of these, five function-switching substitutions (denoted F) eliminated the response to other hormones by repositioning a helix (H7) along one side of the binding cavity and establishing new cortisol-specific contacts. Introducing the F substitutions into AncGR1, however, rendered the protein non-functional (Fig. 1b). The remaining two historical substitutions (P) were permissive: they had no detectable effect on receptor function when introduced into AncGR1, but they allowed F to be tolerated, yielding a cortisol-specific receptor (Fig. 1b). Contingency is apparent, because selection for cortisol specificity could not deterministically drive the acquisition of P, which was required for the subsequent evolution of F and the domain's derived structure and function. It is unlikely that the evolving GR passed through a non-functional intermediate containing F without P[20], because the LBD remained conserved, presumably as a result of functional constraints, for ~40 million years from the gene duplication event that generated it until the evolution of its new function (see ref. 21).

To understand the strength of contingency, we used directed evolution to estimate the frequency of alternative permissive mutations (P′) in a large library of ancestral protein variants. Permissive mutations must fulfil two criteria (Fig. 1c): they must rescue the non-functional AncGR1 +F protein, allowing it to tolerate the F mutations, and they must be compatible with the ancestral sequence and function when introduced into AncGR1. To screen for rescuing mutations that meet the first criterion, we generated a large library of random mutants of AncGR1+F and characterized the resulting distribution of amino acid replacements (Extended Data Figs 1 and 2). We screened this library with a yeast two-hybrid system that linked growth to the cortisol-dependent interaction of the LBD with its coactivator peptide[22,23]. We applied a liberal standard of growth to capture all rescuing mutations and verified their effects

[1]Institute of Molecular Biology and Department of Chemistry & Biochemistry, University of Oregon, Eugene, Oregon 97403, USA. [2]Departments of Human Genetics and Ecology & Evolution, University of Chicago, Chicago, Illinois 60637, USA. [3]Institute of Ecology and Evolution, University of Oregon, Eugene, Oregon 97403, USA.

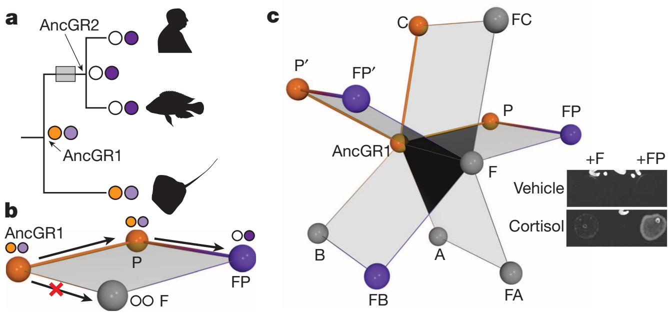


Figure 1 | Searching for alternative permissive mutations in an ancestor of GR. a, Evolution of hormone specificity in vertebrate GRs[6]. Icons indicate taxa (tetrapods, teleosts, elasmobranchs); circles show sensitivity to cortisol (purple) or 11-deoxycorticosterone (orange). The transparent box represents the evolution of new function. **b,** Seven historical substitutions recapitulate the shift in specificity. Two permissive mutations (P), which have no effect on specificity when introduced alone, allow AncGR1 to tolerate five function-switching mutations (F)[6]. Spheres are coloured by primary ligand (orange, 11-deoxycorticosterone; purple, cortisol), or no activation (grey). Thick bars connect functional proteins; thin bars lead to non-functional proteins. Arrows represent evolutionary paths that pass only through functional intermediates. **c,** Historical (P) or alternative permissive (P′) mutations rescue AncGR1+F and are tolerated in the ancestral background. Non-permissive pathways pass through non-functional intermediates (A and B, grey spheres) or fail to rescue F (C). Inset: screening conditions in yeast that identify AncGR1+F variants that confer growth in 1 μM cortisol, compared with vehicle-only control.

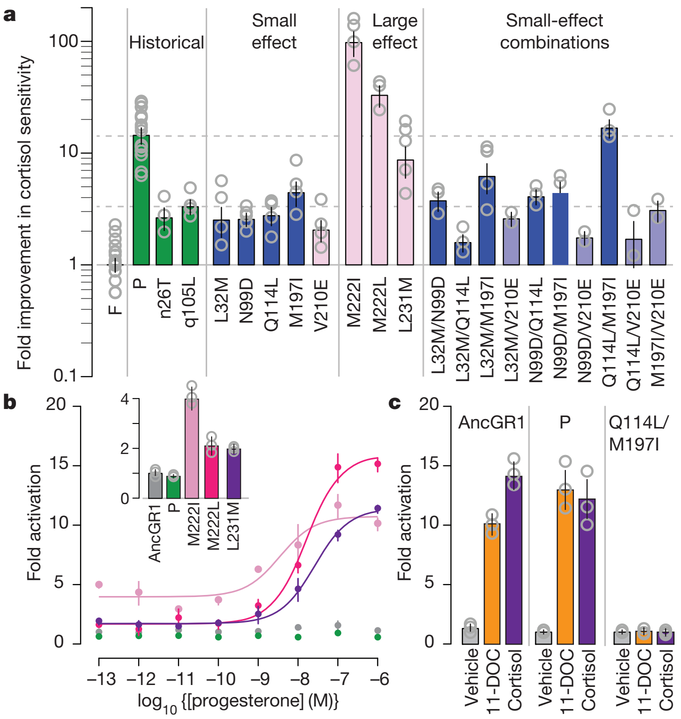


Figure 2 | Rescuing mutations disrupt the ancestral protein's function. a, Effects of rescuing mutations on cortisol sensitivity in AncGR1+F. Sensitivity is defined as the ratio of the mutant to the AncGR1+F concentrations giving half-maximal response (EC_{50}) in a luciferase reporter assay. Results are shown as means and s.e.m. for the number of experimental replicates indicated by grey circles. Green, historical P substitutions, with effect shown by dotted line; rescuing mutations from the screen are coloured by their structural location (see Fig. 3c). **b,** Rescuing AF-H mutations disrupt AncGR1 regulation. Fold reporter activation with progesterone over vector-only control is shown for AncGR1 (grey), historical P (green) and 3 AF-H mutations (pink shades, corresponding to inset graph). Results are shown as means and s.e.m. for three technical replicates. Inset, fold activation for mutants with no hormone (vehicle only). **c,** Q114L/M197I abolishes activation. The fold activation in 1 μM 11-deoxycorticosterone (11-DOC) or cortisol versus vehicle is shown. Results are means ± s.e.m. for three technical replicates.

in both naive yeast and a mammalian reporter assay (Fig. 1c and Extended Data Fig. 3). We screened ~12,500 clones, comprising an estimated 1,025 unique single replacements (71% of all accessible neighbours), 1,802 unique double replacements and 825 higher-order combinations (3,650 total; see Methods and Extended Data Figs 1 and 2); the remainder were duplicate clones or contained nonsense, frameshift or zero non-synonymous mutations. We found no evidence of bias in the library (Extended Data Fig. 2 and Methods).

This screen identified 12 unique clones that improved AncGR1+F's sensitivity to cortisol. These clones carried one, two or three mutations each, but dissection of the combinations showed that functional effects were due entirely to single mutations that co-occurred with neutral changes (Extended Data Fig. 4). In total, we found ten unique single mutations that completely or partly rescued cortisol sensitivity. Two of these involved historically substituted residues: one was a historical P substitution (n26T, with upper and lower cases denoting derived and ancestral states), and the other reverted one F substitution to its ancestral state (I98f), conferring partial growth in the absence of permissive mutations (Extended Data Fig. 3). Of the novel rescuing mutations, three (M222I, M222L and L231M) improved the cortisol-sensitivity of AncGR1+F tenfold or more, an effect as great as historical P (Fig. 2a). The remaining five mutations improved cortisol sensitivity twofold to threefold each, comparable to the individual members of P, but much less than the pair together (Fig. 2a). To see whether pairing any of the small-effect substitutions could recapitulate the effect of P, we generated all twofold combinations of the weak rescuing mutations. Only one pair (Q114L/M197I) affected cortisol sensitivity similarly to the historical set P (Fig. 2a). The screen therefore recovered four alternative rescuing combinations—one double and three single mutants—indicating that rescuing mutations are rare, on the order of 4 in 3,650, or ~0.1%.

To determine whether the rescuing mutations discovered in the screen met the second criterion for permissive mutations—functional compatibility with the ancestral genetic background—we introduced them into AncGR1 and characterized their effects on hormone-dependent activation. Unlike the historical permissive mutations, all four rescuing mutations disrupted the ancestral protein's ligand-regulated transcriptional function. The large-effect rescuing mutations each caused transcriptional activation even in the absence of hormone, and caused promiscuous activation in response to low doses of other steroids such as progesterone (Fig. 2b), a natural hormone excluded by all known extant and ancestral corticosteroid receptors. The pair Q114L/M197I destroyed AncGR1's transcriptional function entirely, making it unable to activate reporter expression even at high hormone concentrations (Fig. 2c).

Permissive mutations are therefore extremely rare. Among ~3,660 unique protein variants (~3,650 in the screened library plus 10 engineered double mutants), zero permissive genotypes were present. One permissive combination, the historical set P, exists in the universe of sequences near AncGR1, so we estimate an upper bound frequency of accessible permissive pathways of less than 1 in 3,660 (0.03%). The total frequency is probably far lower, because knowledge of this one permissive pathway was not acquired by sampling. Further, our screen of double mutants was biased towards the discovery of rescuing variants, because it included engineered combinations of all single mutations that had a detectable rescuing effect. The universe of possible variants containing two or more replacements is very large, so alternative permissive sets may exist; however, these genotypes would require multiple independent substitutions, and the joint probability of such events would be very low because they cannot be acquired deterministically by selection for the derived function. A permissive mutation might conceivably be subject to selection for some other function; however, unless the selected and derived functions are correlated, the probability that selection would deterministically fix a compound permissive genotype is extremely low.

Evolution of the F mutations was therefore strongly contingent on prior low-probability events.

To understand the mechanisms that make permissive mutations both necessary and rare, we characterized the biophysical effects of F, P, and the four sets of rescuing but non-permissive mutations. Permissive mutations are often thought to act through effects on the global stability of folding: function-switching mutations destabilize a protein, making it prone to degradation and aggregation, but permissive mutations increase stability, and offset this effect[13,15,24,25]. Structural considerations suggested that a stability tradeoff might explain the effects of F and P. The F mutations cause a 3 Å shift in the position of helix H7 relative to H10 and the ligand, disrupting numerous contacts; they also open empty space between the ligand and helix H3 and remove a hydrogen bond from the key loop that connects AF-H to H10 (refs 21, 26). In contrast, the P mutations add favourable interactions—both a new hydrogen bond and improved packing interactions—in the crystal structure and in molecular dynamics (MD) simulations (Extended Data Fig. 5). To elucidate the effects of F and P on stability, we measured the midpoint of irreversible thermal denaturation (T_m) of steroid-bound AncGR1 containing each of the historical F and P mutations. As expected, each F mutation except l111Q was destabilizing (Extended Data Fig. 6a), and the P mutations were stabilizing (Fig. 3a).

Although these data are consistent with the global stability model, several other observations are inconsistent with it. First, the F and P mutations did not affect expression in mammalian cells as measured by western blot analysis (Extended Data Fig. 6b), indicating that AncGR1-F is functionally compromised rather than subject to degradation or aggregation because of reduced stability. Second, under the global stability model, rescuing mutations should be more frequent than we observed. The global model predicts that any stabilizing mutation should be permissive[24,25], and it is estimated that 1–10% of mutations are stabilizing[27]; however, only ~0.1% of our library was rescuing, and permissive mutations were even rarer. Third, the global stability model predicts that any rescuing mutation should also be permissive, but we found that several rescuing mutations were deleterious when introduced without the function-switching mutations. Finally, the rescuing mutants all increased the T_m of AncGR1+F more than they did in AncGR1, suggesting a specific epistatic effect rather than a generic compensatory mechanism (Fig. 3b and Extended Data Table 1). These observations all indicate that permissive mutations must do more than simply increase global stability.

To understand the requirements that permissive mutations must fulfil, we first examined the location of permissive and rescuing mutations in the protein's structure. Under the global stability model, a stabilizing mutation should be permissive, irrespective of its location[24,28]. In contrast, the permissive and rescuing mutations exhibited a striking structural distribution, occurring in two distinct clusters near the F mutations: 'pocket' substitutions bordering the ligand cavity, and 'AF-H' substitutions at the interface between AF-H and the rest of the protein (Fig. 3c). Both the ancestral crystal structures and MD simulations showed that the historical P mutations yielded new favourable contacts involving the same structural elements destabilized by F (Extended Data Fig. 5). Specifically, Thr 26 strengthens a hydrogen bond connecting helix H3 to the H10/AF-H loop, compensating for the loss of a hydrogen bond in this loop as a result of F mutation s212Δ. Leu 105 improves packing interactions between helices H3 and H7, apparently compensating for the effects of the other F mutations on the interactions between H3, H7 and the ligand. Similarly, all rescuing mutations we discovered in our screen improved packing interactions involving AF-H or H7 (Fig. 4 and Extended Data Figs 7 and 8).

These observations suggest that permissive mutations must stabilize specific local structural elements destabilized by F, rather than generically modulating global stability. To test this hypothesis, we used the structure to identify a potentially stabilizing pair of mutations (E165A and K168E) ~25 Å distant from the ligand pocket and AF-H (Fig. 3c). We introduced them into AncGR1+F and found that they raised T_m by 1.4 °C; rather than rescuing function, however, they impaired AncGR1+F's cortisol sensitivity roughly tenfold (Fig. 3d). These data confirm that increasing global stability is not sufficient to yield a permissive effect and point to a biophysical requirement that limits the number of permissive mutations: they must exert specific local rather than generic global effects on protein stability.

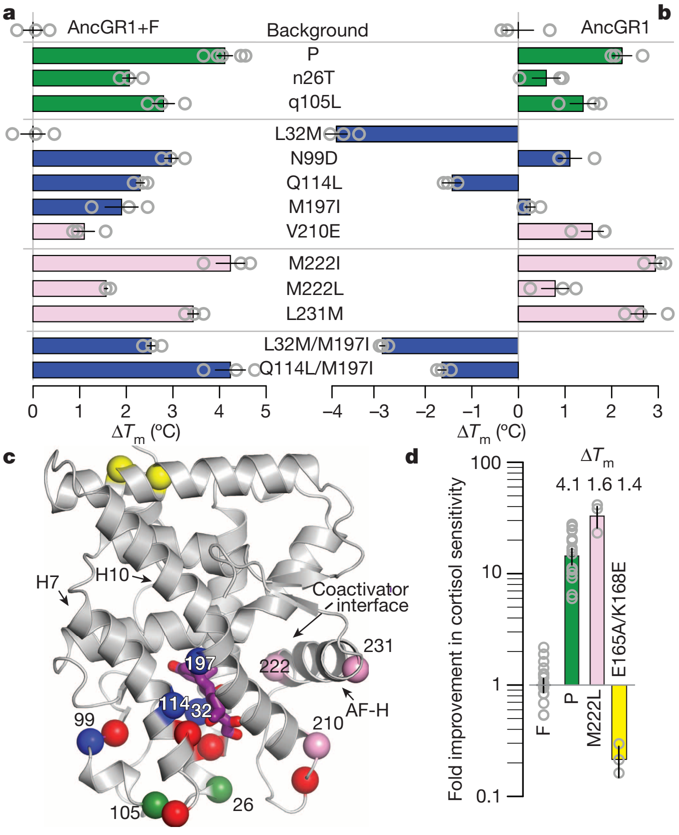


Figure 3 | Permissive mutations must stabilize local structural elements. **a**, **b**, Effect of rescuing mutations on T_m values of AncGR1+F (**a**) and AncGR1 (**b**). Colours correspond to structural position in **c**. **c**, Structural distribution of mutations on AncGR1 (PDB 3RY9). Spheres, Cα atoms. Red, historical F substitutions; green, historical P; blue, rescuing ligand-pocket mutations; pink, rescuing AF-H mutations; yellow, distant mutations that stabilize but do not rescue. Purple sticks show cortisol; helices are indicated. **d**, Change in cortisol sensitivity caused by E165A/K168E in AncGR1+F (yellow bar). Effects of P and M222L are shown for comparison. ΔT_m values relative to AncGR1+F are shown. Results in **a**, **b** and **d** are shown as means and s.e.m. for the experimental replicates indicated by grey circles.

This requirement explains why rescuing mutations were few, but it does not explain why they were functionally incompatible with AncGR1, suggesting that further biophysical requirements limit the number of permissive mutations. To elucidate these requirements, we first examined the mechanisms by which the large-effect rescuing mutations make the ancestral protein super-active. All three increased the stability of both AncGR1 and AncGR1+F (Fig. 3a, b) and are clustered on AF-H, suggesting that they exert their effect by disrupting ligand-induced allosteric regulation of this helix's position (Fig. 3c), which differentiates inactive and active conformations. For a properly regulated receptor without ligand, the inactive conformation is more stable than the active conformation and thus the dominant species (Fig. 4a); binding of hormone stabilizes the active conformation, causing it to become dominant. To test whether the AF-H mutations unconditionally stabilized the active conformation, we performed MD simulations of these mutations in AncGR1 in the absence of ligand. As predicted, M222I and M222L improved hydrophobic packing between the active position of AF-H and helix H3 (Fig. 4b, c), and L231M introduced a new sulphur–π interaction, anchoring AF-H in the active position against H10 (Extended

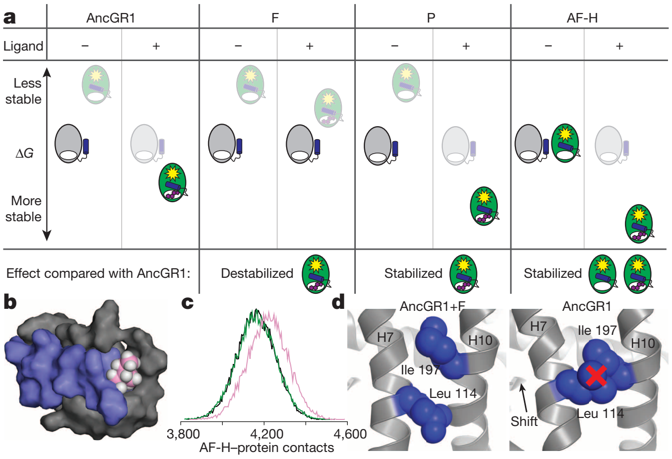


Figure 4 | Biophysical requirements make some rescuing mutations intolerable in the ancestral protein. a, A simple thermodynamic model explains why AF-H mutants lead to activity in the absence of hormone. The protein can exist in inactive (grey) or active (green) microstates, which are differentiated by AF-H's position (blue). For each genotype, the relative free energy (Δ*G*) of active and inactive states is shown with or without hormone. Populated states are opaque; unpopulated states are faded. **b**, Snapshot from MD trajectory of AncGR1+M222I shows tight packing interaction between Ile 222 (pink) and the rest of the protein. Blue, AF-H; grey, surface that AF-H contacts. **c**, Distribution of atom contacts (centre-to-centre distances 3.5 Å or less) between AF-H and the rest of the protein over three replicate MD trajectories for AncGR1+F (black), +P (green) and +M222I (pink). The *y* axis is frequency. **d**, Change in position of H7 with respect to H10 from ancestral to derived GRs changes the effects of mutations Q114L/M197I from incompatible to rescuing (blue spheres). Structures are AncGR2 (left, PDB 3GN8) and AncGR1 (right, PDB 3RY9) with side chains at these sites introduced (spheres).

Data Fig. 7). Stabilizing the active conformation relative to the inactive conformation is expected to increase the proportion of the protein in the active conformation, explaining why these mutations imparted activity in the absence of ligand and made the receptor highly sensitive to formerly weak ligands (Fig. 4a). These observations point to a second limiting requirement: permissive mutations must not alter the energetic balance between functional conformations of the protein. That is, they must stabilize the 'right' portions of the protein without stabilizing the 'wrong' portion. The global stability model does not account for these constraints because GR function depends not only on the stability of folded versus unfolded or misfolded forms but also on the stabilities of active versus inactive conformations in both the presence and the absence of ligand.

Finally, we examined why the rescuing pair Q114L/M197I rendered the ancestral protein non-functional (Fig. 2c). These sites are near the ligand-binding pocket, facing each other on helices H7 and H10 (Fig. 4d). In the presence of F, the two residues are slightly offset, and the rescuing states Leu 114 and Ile 197 improve hydrophobic packing between H7 and H10, explaining their observed positive effect on the derived protein's stability and sensitivity (Extended Data Fig. 8). In the AncGR1 structure, however, the shifted position of H7 places these two residues directly across from each other: the large side chains of the rescuing residues clash and destabilize the H7/H10 interaction (Fig. 4d). As predicted by this model, the pair of rescuing states increases the T_m of AncGR1+F but lowers that of AncGR1 (Fig. 3b). These observations reveal a final requirement: permissive mutations must be compatible with the conformations of both the ancestral and derived proteins.

Evolutionary contingency has usually been discussed in terms of chance external forces, such as random extinction by asteroid impacts or climate change[2]. Our results show that the internal organization of biological systems—in this case, a protein's structure and thermodynamics—can give rise to strong contingency during evolution. The F mutations that triggered GR's functional transition required permissive mutations to stabilize the specific local structural elements that F destabilized, without disturbing the energetic balance between the receptor's functional conformations or clashing with ancestral or derived protein structures. Our data indicate that very few mutations can satisfy all these biophysical requirements, making GR's evolution dependent on rare, low-probability historical events.

Our findings point to strong contingency in the evolution both of GR's primary sequence and of its molecular form—the structural and mechanistic underpinnings that produce the protein's function. GR's cortisol specificity was achieved by a unique repositioning of H7 and the reorganization of numerous hormone contacts. If other F-like mutations exist that could produce a form and function similar to those of the modern GR, these mutations would reorganize and destabilize the same local elements of the ligand–receptor complex. To be tolerated, these effects would have to be offset by permissive mutations. The permissive mutations, in turn, would be subject to the same biophysical constraints as the historical permissive mutations, because those constraints arise from the functional form itself and the fundamental architecture of the GR LBD. Our experiments establish that very few accessible genotypes satisfy these constraints. Permissive sequence changes that could enable alternative ways of achieving a similar form and function—even using entirely different mutations–would therefore also be very rare.

If evolutionary history could be replayed from the ancestral starting point, the same kind of permissive substitutions would be unlikely to occur. The transition to GR's present form and function would probably be inaccessible, and different outcomes would almost certainly ensue. Cortisol-specific signalling might evolve by a different mechanism in the GR, or by an entirely different protein, or not at all; in each case, GR—or the vertebrate endocrine system more generally—would be substantially different. Because GR is the only ancestral protein for which alternative evolutionary trajectories to historically derived functions have been explored, the generality of our findings is unknown. The specific biophysical constraints, and in turn the degree and nature of contingency, that shape the evolution of other proteins are likely to depend on the particular architecture of each protein and the unique historical mechanisms by which its functions evolved.

METHODS SUMMARY

The AncGR1+F mutant library was generated by GeneMorphII-EZClone, using conditions to maximize single and double mutations. We characterized the library's composition by sequencing random clones. For yeast two-hybrid screening, we cloned the LBD library into pBDGAL4 and the human steroid receptor coactivator peptide SRC-1 into pADGAL4 (refs 22, 23). Clones showing any growth in the screen were retransformed into naive yeast and characterized for hormone-dependent growth, then subcloned into pSG5-DBD, transfected into CHO-K1 cells, and assayed using a dual-luciferase reporter. Additional genotypes were generated by Quikchange mutagenesis. Proteins were expressed as His-tagged fusions with maltose-binding protein (MBP), then cleaved and purified to more than 99% purity by sequential affinity chromatography. We followed irreversible thermal denaturation using circular dichroism. For MD simulations, three independent 100-ns trajectories were performed for each genotype, using GROMACS 4.5.5 and the CHARMM27 force field starting from equilibrated crystallographic coordinates with or without *in silico* mutations.

Online Content Any additional Methods, Extended Data display items and Source Data are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 24 January; accepted 28 April 2014.
Published online 15 June 2014.

1. Monod, J. *Chance and Necessity: An Essay on the Natural Philosophy of Modern Biology* (Vintage Books, 1972).
2. Gould, S. J. *Wonderful Life: The Burgess Shale and the Nature of History* (W. W. Norton & Company, 1990).
3. Losos, J. B., Jackman, T. R., Larson, A., Queiroz, K. & Rodríguez-Schettino, L. Contingency and determinism in replicated adaptive radiations of island lizards. *Science* **279,** 2115–2118 (1998).
4. Morris, S. C. *The Crucible of Creation: The Burgess Shale and the Rise of Animals* (Oxford Univ. Press, 2000).
5. Beatty, J. Replaying life's tape. *J. Phil.* **103,** 336–362 (2006).

6. Ortlund, E. A., Bridgham, J. T., Redinbo, M. R. & Thornton, J. W. Crystal structure of an ancient protein: evolution by conformational epistasis. *Science* **317,** 1544–1548 (2007).
7. Blount, Z. D., Borland, C. Z. & Lenski, R. E. Historical contingency and the evolution of a key innovation in an experimental population of *Escherichia coli. Proc. Natl Acad. Sci. USA* **105,** 7899–7906 (2008).
8. Travisano, M., Mongold, J. A., Bennett, A. F. & Lenski, R. E. Experimental tests of the roles of adaptation, chance, and history in evolution. *Science* **267,** 87–90 (1995).
9. Meyer, J. R. *et al.* Repeatability and contingency in the evolution of a key innovation in phage lambda. *Science* **335,** 428–432 (2012).
10. Fisher, R. A. *The Genetical Theory of Natural Selection* (Oxford Univ. Press, 1958).
11. Martin, R. E. *et al.* Chloroquine transport via the malaria parasite's chloroquine resistance transporter. *Science* **325,** 1680–1682 (2009).
12. Field, S. F. & Matz, M. V. Retracing evolution of red fluorescence in GFP-like proteins from faviina corals. *Mol. Biol. Evol.* **27,** 225–233 (2010).
13. Bloom, J. D., Gong, L. I. & Baltimore, D. Permissive secondary mutations enable the evolution of influenza oseltamivir resistance. *Science* **328,** 1272–1275 (2010).
14. Lynch, V. J., May, G. & Wagner, G. P. Regulatory evolution through divergence of a phosphoswitch in the transcription factor CEBPB. *Nature* **480,** 383–386 (2011).
15. Gong, L. I., Suchard, M. A. & Bloom, J. D. Stability-mediated epistasis constrains the evolution of an influenza protein. *eLife* **2,** e00631 (2013).
16. Peisajovich, S. G. & Tawfik, D. S. Protein engineers turned evolutionists. *Nature Methods* **4,** 991–994 (2007).
17. Romero, P. A. & Arnold, F. H. Exploring protein fitness landscapes by directed evolution. *Nature Rev. Mol. Cell Biol.* **10,** 866–876 (2009).
18. Bledsoe, R. K., Stewart, E. L. & Pearce, K. H. Structure and function of the glucocorticoid receptor ligand binding domain. *Vitamins Hormones* **68,** 49–91 (2004).
19. Moras, D. & Gronemeyer, H. The nuclear receptor ligand-binding domain: structure and function. *Curr. Opin. Cell Biol.* **10,** 384–391 (1998).
20. Smith, J. M. Natural selection and the concept of a protein space. *Nature* **225,** 563–564 (1970).
21. Carroll, S. M., Ortlund, E. A. & Thornton, J. W. Mechanisms for the evolution of a derived function in the ancestral glucocorticoid receptor. *PLoS Genet.* **7,** e1002117 (2011).
22. Ding, X. F. *et al.* Nuclear receptor-binding sites of coactivators glucocorticoid receptor interacting protein 1 (GRIP1) and steroid receptor coactivator 1 (SRC-1): multiple motifs with different binding specificities. *Mol. Endocrinol.* **12,** 302–313 (1998).
23. Chen, Z., Katzenellenbogen, B. S., Katzenellenbogen, J. A. & Zhao, H. Directed evolution of human estrogen receptor variants with significantly enhanced androgen specificity and affinity. *J. Biol. Chem.* **279,** 33855–33864 (2004).
24. Bershtein, S., Segal, M., Bekerman, R., Tokuriki, N. & Tawfik, D. S. Robustness–epistasis link shapes the fitness landscape of a randomly drifting protein. *Nature* **444,** 929–932 (2006).
25. Bloom, J. D., Labthavikul, S. T., Otey, C. R. & Arnold, F. H. Protein stability promotes evolvability. *Proc. Natl Acad. Sci. USA* **103,** 5869–5874 (2006).
26. Bridgham, J. T., Ortlund, E. A. & Thornton, J. W. An epistatic ratchet constrains the direction of glucocorticoid receptor evolution. *Nature* **461,** 515–519 (2009).
27. Tokuriki, N., Stricher, F., Schymkowitz, J., Serrano, L. & Tawfik, D. S. The stability effects of protein mutations appear to be universally distributed. *J. Mol. Biol.* **369,** 1318–1332 (2007).
28. Bloom, J. D., Arnold, F. H. & Wilke, C. O. Breaking proteins with mutations: threads and thresholds in evolution. *Mol. Syst. Biol.* **3,** 76 (2007).

Acknowledgements We thank J. Bridgham, members of the Thornton laboratory, and B. Buckley McAllister for technical assistance and fruitful discussions. We thank M. Stallcup for sharing plasmids and the University of Oregon ACISS cluster for computing resources (National Science Foundation (NSF) OCI-0960354). This work was supported by National Institutes of Health (NIH) F32-GM090650 (M.J.H.), NIH R01-GM081592 (J.W.T.) and R01-GM104397 (J.W.T.), NSF IOB-0546906 (J.W.T.) and a Howard Hughes Medical Institute Early Career Scientist award (J.W.T.).

Author Contributions M.J.H. and J.W.T. conceived the project, designed the experiments and wrote the paper. M.J.H. performed the experiments and analysed the data.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to J.W.T. (joet1@uchicago.edu).

doi:10.1038/nature13401

DENR–MCT-1 promotes translation re-initiation downstream of uORFs to control tissue growth

Sibylle Schleich[1,2], Katrin Strassburger[1]*, Philipp Christoph Janiesch[2]*, Tatyana Koledachkina[2], Katharine K. Miller[2], Katharina Haneke[1,3], Yong-Sheng Cheng[1], Katrin Küchler[2], Georg Stoecklin[1,3], Kent E. Duncan[2]* & Aurelio A. Teleman[1]*

During cap-dependent eukaryotic translation initiation, ribosomes scan messenger RNA from the 5′ end to the first AUG start codon with favourable sequence context[1,2]. For many mRNAs this AUG belongs to a short upstream open reading frame (uORF)[3], and translation of the main downstream ORF requires re-initiation, an incompletely understood process[1,4–6]. Re-initiation is thought to involve the same factors as standard initiation[1,5,7]. It is unknown whether any factors specifically affect translation re-initiation without affecting standard cap-dependent translation. Here we uncover the non-canonical initiation factors density regulated protein (DENR) and multiple copies in T-cell lymphoma-1 (MCT-1; also called MCTS1 in humans) as the first selective regulators of eukaryotic re-initiation. mRNAs containing upstream ORFs with strong Kozak sequences selectively require DENR–MCT-1 for their proper translation, yielding a novel class of mRNAs that can be co-regulated and that is enriched for regulatory proteins such as oncogenic kinases. Collectively, our data reveal that cells have a previously unappreciated translational control system with a key role in supporting proliferation and tissue growth.

Cellular protein abundance depends largely on mRNA translation[8]. Little is known about how translation of specific sets of mRNAs can be coordinately regulated[9,10]. mRNAs with uORFs require re-initiation[11–14], whereby ribosomes translate the uORF, terminate and then restart translating the main ORF[1,4,6,15]. No metazoan *trans*-acting factors have yet been described that selectively affect re-initiation, enabling coordinate regulation of uORF-containing mRNAs.

eIF2D (also called ligatin) and the related DENR–MCT-1 complex are candidate re-initiation regulators. They associate with 40S ribosomal subunits and have domains implicated in RNA binding and start codon recognition (Extended Data Fig. 1a). *In vitro* they can recycle post-termination complexes, recruit initiator methionyl-tRNA (Met-tRNA$_i^{Met}$) to mRNAs containing viral internal ribosome entry sites[16,17], and affect movement of post-termination 80S complexes to nearby AUG codons[6]. DENR–MCT-1 has not previously been implicated in re-initiation, and MCT-1 is an oncogene affecting cellular mRNA translation by an unclear mechanism[18–20]. Collectively, these studies suggest that DENR–MCT-1 and eIF2D might regulate translation of cancer-relevant mRNAs through non-canonical mechanisms.

To study DENR function, we generated *Drosophila* knockouts for the homologous gene (CG9099), lacking DENR transcript or protein (DENR^{KO}, Extended Data Fig. 1b–d). DENR^{KO} flies die as pharate adults with a larval-like epidermis (Fig. 1a), due to impaired proliferation of histoblast cells (Fig. 1b and Extended Data Fig. 1e). This is rescued by expressing DENR ubiquitously (Tubulin–GAL4) or specifically in histoblast cells (Escargot–GAL4) (χ^2 test $P < 0.05$, Extended Data Fig. 1f). Although DENR is expressed ubiquitously (Extended Data Fig. 1g), quickly proliferating histoblast cells appear more sensitive to DENR loss than non-proliferating tissues. DENR^{KO} flies also have crooked legs and incorrectly rotated genitals (Fig. 1c and Extended Data Fig. 1h–h′). These phenotypes are not observed in mutants with generally impaired translation (*Minutes*; ref. 21), but are found in flies with reduced cell-cycle regulators or Ecdysone Receptor signalling[22,23], suggesting that DENR affects translation of a subset of mRNAs involved in cell proliferation and signalling.

Similar phenotypes were observed in flies expressing RNA interference (RNAi) targeting MCT-1 (fly homologue CG5941; Extended Data Fig. 1i), which like human MCT-1 binds DENR (Extended Data Fig. 1j). Reducing *ligatin* (fly homologue of human eIF2D) gene dosage in DENR^{KO} flies caused fewer animals to reach pupation (χ^2 test $P < 0.05$, Extended Data Fig. 1k), indicating that DENR^{KO} phenotypes result from loss of DENR–MCT-1 complex with eIF2D-like activity.

DENR^{KO} larvae and DENR knockdown S2 cells grow slowly with reduced protein accumulation rates (Fig. 1d, e, g). Mutant polysome profiles show reduced polysome/monosome ratios (Fig. 1f, h and Extended Data Fig. 1l–n′), suggesting defective translation initiation. Despite more ribosomes and initiator tRNA per cell (Extended Data Fig. 1o, p), DENR knockdown cells have reduced protein synthesis rates when proliferating (Fig. 1i). When quiescent, DENR knockdown cells no longer display these phenotypes, and become enlarged compared to controls (Fig. 1h, i (right panel) and Extended Data Fig. 1q–r). Thus, DENR promotes translation of cellular mRNAs in proliferating but not quiescent cells.

We identified ~100 mRNAs requiring DENR for efficient translation by profiling actively translated mRNAs from 80S and polysome fractions of control and DENR knockdown cells and normalizing to total mRNA (Supplementary Table 1). We further analysed *myoblast city* (*mbc*) because it was the second most under-translated mRNA and we could obtain antibody to detect it. Quantitative polymerase chain reaction with reverse transcription (RT–PCR) confirmed that *mbc* mRNA is under-represented in polysomes of DENR knockdown cells (Fig. 2a), leading to reduced Mbc protein but not mRNA (Fig. 2b), whereas other proteins were not reduced (Extended Data Fig. 2a). The *mbc* 5′ UTR was sufficient to impart DENR-dependence to a *Renilla* luciferase (RLuc) reporter (Extended Data Fig. 2b). This DENR dependence requires the 5′ cap and is not accompanied by a drop in general translation (Extended Data Fig. 2c, d′ and Fig. 2c). Combined knockdown of DENR and MCT-1 had no additive effect, as they are a functional complex (Extended Data Fig. 2e). In sum, the DENR–MCT-1 complex selectively promotes cap-dependent translation of *mbc* via its 5′ UTR.

Systematic 5′ UTR truncations (Extended Data Fig. 2f–h) identified 175 nucleotides necessary and sufficient for DENR dependence (Fig. 2d and Extended Data Fig. 3a, b) containing 3 uORFs with strong Kozak sequences (stuORFs, red boxes in Fig. 2d). Mutating all three stuORF ATGs, or their Kozak sequences, abolished DENR dependence (Fig. 2e, f and Extended Data Fig. 3c), indicating that translation initiation on these stuORFs is necessary for DENR dependence. No additional *cis*-acting sequences were necessary; removing sequences upstream, downstream, or between the uORFs, or mutating the uORF coding sequences, did not affect DENR dependence (Fig. 2g and Extended Data Fig. 2f–h). Two possible explanations are: (1) DENR promotes bypass of stuORF initiation codons; and (2) DENR affects re-initiation after stuORF translation.

[1]German Cancer Research Center (DKFZ), 69120 Heidelberg, Germany. [2]Center for Molecular Neurobiology (ZMNH), University Medical Center Hamburg-Eppendorf (UKE), Falkenried 94, 20251 Hamburg, Germany. [3]Zentrum für Molekulare Biologie der Universität Heidelberg (ZMBH), DKFZ-ZMBH Alliance, 69120 Heidelberg, Germany.
*These authors contributed equally to this work.

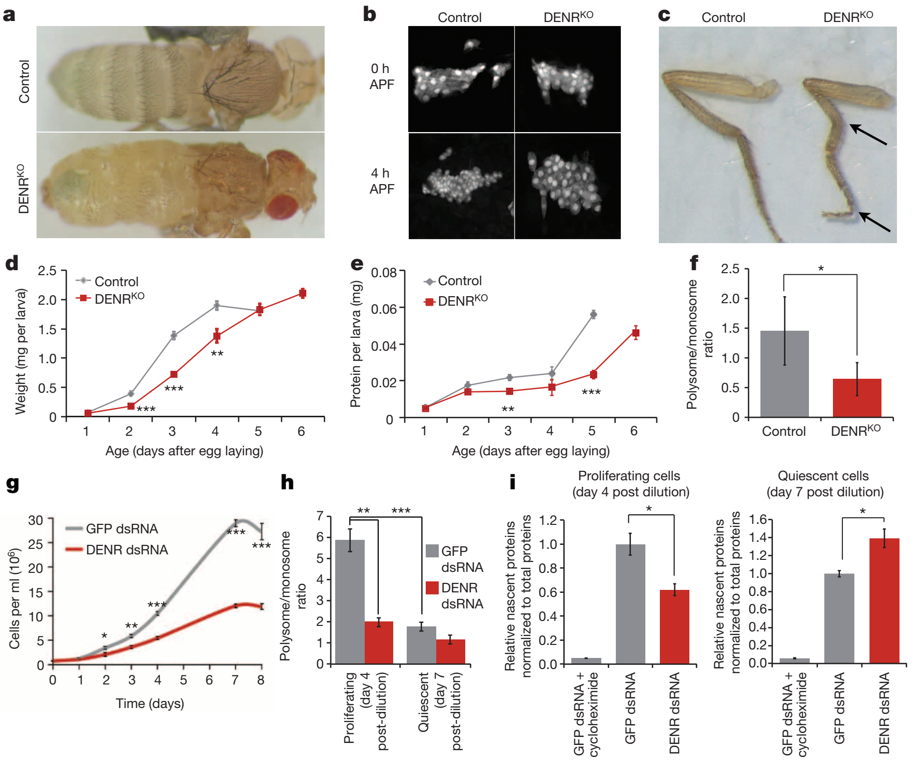


Figure 1 | DENR promotes cell proliferation and boosts protein synthesis in proliferating but not quiescent cells. **a**, DENR^{KO} flies die as pharate adults with larval-like abdominal epidermis. **b**, DENR^{KO} anterior–dorsal histoblast nests have correct cell numbers at onset of pupation (0 h after puparium formation, APF), but impaired proliferation during the first 4 h of pupal development (4 h APF) (visualized by esgG4>GFP). **c**, DENR^{KO} flies have crooked legs. **d**, **e**, DENR^{KO} flies accrue mass (**d**) and protein (**e**) slowly, pupating with 1 day delay (day 6 data point, absent in controls). **f**, Polysome profiles from DENR^{KO} larvae have reduced polysome/monosome ratios. **g**, DENR knockdown cells proliferate slowly and enter quiescence at low cell density. **h**, Proliferating but not quiescent polysome profiles of DENR knockdown cells have reduced polysome/monosome ratios. **i**, Proliferating (left) but not quiescent (right) DENR knockdown cells have reduced *de novo* protein synthesis rates, quantified by metabolic labelling with methionine analogue. Error bars indicate standard deviation (s.d.) (**d**–**f**) or s.e.m. (**g**–**i**). *t*-test (**d**–**f**) or Mann–Whitney *U*-test (**g**–**i**′) $*P < 0.05$, $**P < 0.01$, $***P < 0.001$.

In model 1, the stuORF stop codon is irrelevant because DENR would act at the stuORF start codon. In model 2 the stuORF stop codon is crucial as translation re-initiation on the main ORF only occurs after termination on the stuORF. Two point mutations removing the stuORF stop codons completely abolished DENR dependence (Fig. 2h and Extended Data Fig. 3d), indicating that DENR promotes translation re-initiation.

Introducing synthetic stuORFs into a control reporter was sufficient to impart DENR dependence, with multiple stuORFs acting additively (Fig. 3a). Re-initiation efficiency is reportedly inversely related to uORF length, presumably because initiation factors dissociate from ribosomes as elongation proceeds[7]. Consistently, the ability of DENR to promote re-initiation dropped as uORFs became longer (Fig. 3a, right panel), reaching

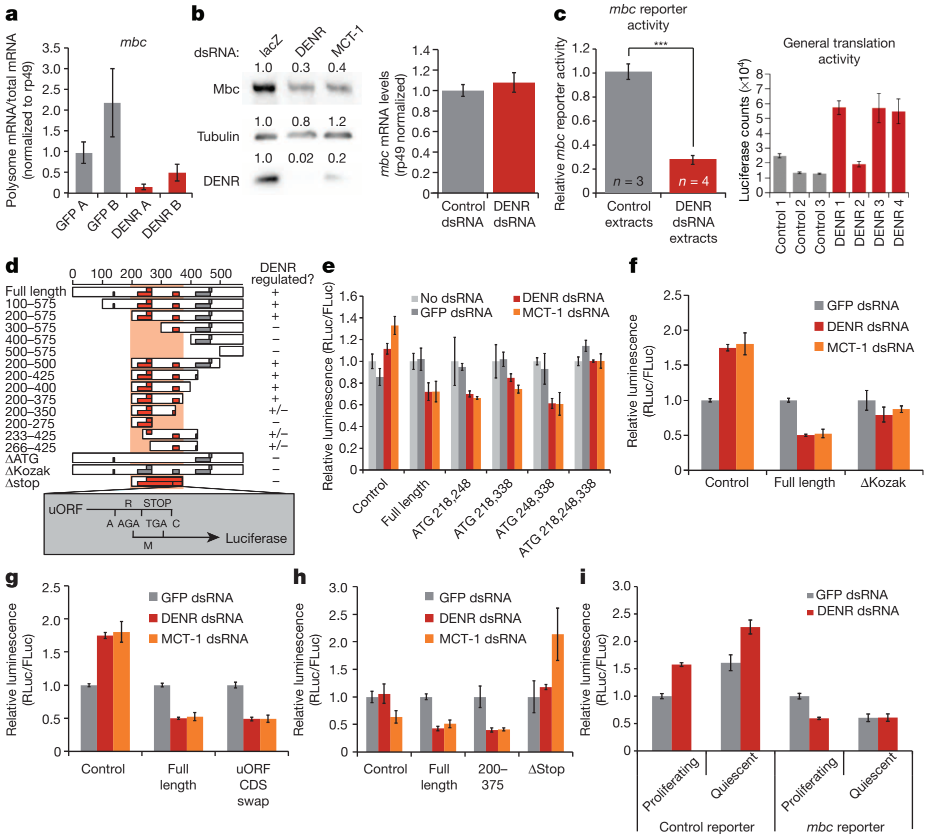


Figure 2 | DENR promotes re-initiation of translation downstream of uORFs in the *mbc* 5′ UTR. **a**, qRT–PCR validation that the *mbc* mRNA is preferentially depleted from polysomes in DENR knockdown cells (DENR A and B) compared to controls (GFP A and B). **b**, Mbc protein (left) but not mRNA (right) levels are reduced in DENR and MCT-1 knockdown cells. **c**, Translation extracts from DENR knockdown cells are impaired in translating a reporter containing the *mbc* 5′ UTR (left) but not in translating a control RLuc reporter mRNA without uORFs in the 5′ UTR (right). **d**, Schematic overview of the *mbc* 5′ UTR and the tested DNA reporter constructs, summarizing results from other panels as well as multiple (≥3) additional replicates on all the luciferase assays, not shown. Details in Extended Data Fig. 3. **e**, **f**, Mutating the start codons of the three *mbc* uORFs with strong Kozak sequences (**e**) or their Kozak sequences to the less functional gtgtATG (**f**) blunts regulation by DENR. **g**, Mutating the coding sequence of *mbc* uORFs to poly-glutamine has no effect on DENR regulation. **h**, Mutation of the stop codons of *mbc* uORFs 218, 248 and 338, as diagrammatically shown in **d**, causing the uORFs to extend past the RLuc ATG, leads to loss of DENR-dependent regulation. **i**, DENR knockdown leads to impaired expression of the *mbc* 5′ UTR RLuc reporter in proliferating but not quiescent S2 cells. Error bars: s.d. *t*-test $*P < 0.05$, $***P < 0.001$.

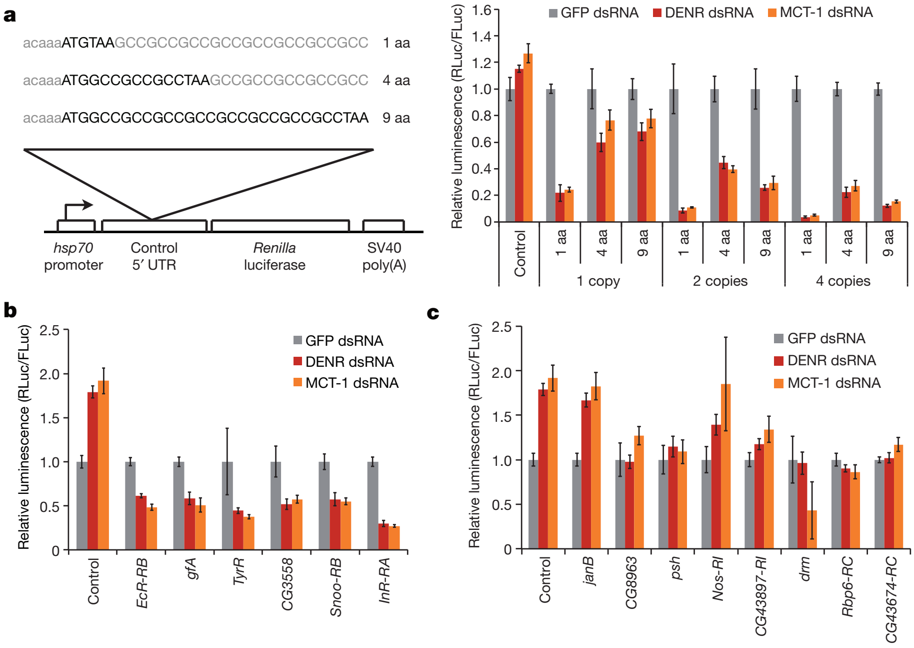


Figure 3 | uORFs with strong Kozak sequences (stuORFs) are sufficient to impart DENR–MCT-1-dependent regulation. **a**, Introduction of synthetic uORFs bearing a 'strong' Kozak into a control 5′ UTR imparts DENR-dependent regulation (DNA reporters). **b**, **c**, 5′ UTRs bearing stuORFs are all DENR-dependent (**b**) whereas 5′ UTRs lacking uORFs (**c**) are not (DNA reporters). Error bars indicate s.d.

zero effect on a dicistronic transcript containing a long upstream ORF (not shown). In sum, mRNAs display a continuum of DENR dependence, depending on the number and length of the uORFs and the strength of their Kozak sequences. A computational search revealed thousands of 5′ UTRs containing uORFs (Extended Data Fig. 4a). We generated a predicted 'DENR-dependence score' for all transcripts based on the number of uORFs they contain and the strength of their Kozak sequences (Extended Data Fig. 4b, b′ and Supplementary Table 2). Transcripts with high DENR-dependence scores were significantly enriched among the mRNAs with reduced translation on DENR knockdown (Extended Data Fig. 4c, d), suggesting a general mechanism. We tested ten 5′ UTRs predicted to be DENR-dependent using luciferase assays. Six conferred DENR dependence (Fig. 3b), and four inhibited reporter translation too strongly to test experimentally. Conversely, 16 5′ UTRs without uORFs were not DENR-dependent (Fig. 3c and data not shown). Therefore, 5′ UTRs with stuORFs are DENR-dependent, identifying a new class of transcripts for which translation can be co-regulated. Gene Ontology analysis[24] revealed that these genes are enriched for transcriptional regulators and kinases (Extended Data Fig. 4e, f).

Immunoprecipitation of DENR showed that it binds mRNAs containing or lacking stuORFs (Extended Data Fig. 4g), suggesting that it interacts generally with initiating ribosomes, but is required on stuORF-containing mRNAs. Because only 15% of genes contain stuORFs, we were surprised to see global effects on polysomes upon DENR knockdown (Extended Data Fig. 1l). A DENR knockdown time course revealed that stuORF-dependent translation drops before changes in polysome or ribosome levels (Extended Data Fig. 5), indicating that these are probably secondary consequences.

Because Insulin and Ecdysone receptors (InR and EcR) contain DENR-dependent 5′ UTRs (Fig. 3b), we asked whether impaired InR and EcR translation contribute to DENR^{KO} phenotypes. Loss of DENR–MCT-1 function in S2 cells or DENR^{KO} animals leads to reduced InR and EcR proteins, but not mRNA levels, and reduced InR and EcR signalling (Fig. 4a–c and Extended Data Fig. 6a–c′). Reconstituting InR/EcR expression in DENR^{KO} animals partially but significantly rescued developmental rate and histoblast proliferation (Fig. 4d, e and Extended Data Fig. 6d). Thus, loss of DENR–MCT-1 causes reduced InR/EcR translation and signalling, and consequently impaired cell proliferation and organismal development.

Data from DENR^{KO} flies and DENR knockdown cells suggested that proliferating cells are phenotypically more sensitive to DENR loss-of-function than quiescent cells. One explanation could be that DENR activity is low in quiescent cells, hence its removal has little effect. Using *mbc* and stuORF reporters, and endogenous *mbc* translation, as read-outs for DENR activity revealed that DENR loss had a larger impact in proliferating compared to quiescent cells (Fig. 2i and Extended Data Fig. 7). Hence DENR–MCT-1 present in quiescent cells is not very active.

To study DENR function *in vivo*, we generated flies carrying fluorescent reporters with or without a stuORF (Extended Data Fig. 8a). These reporters have identical promoters, 5′ UTRs and 3′ UTRs, and are integrated in exactly the same genomic locus via phiC31-mediated recombination, ensuring their identical transcription. This revealed that DENR promotes stuORF reporter, but not control reporter, expression in animals (Extended Data Fig. 8). Because stuORF–GFP reporter expression is entirely DENR-dependent, it serves as an *in vivo* DENR activity readout. Interestingly, the larval anterior, which contains proliferating tissues like brain and imaginal discs, shows stronger DENR activity than other larval regions (Extended Data Fig. 8b). Inclusion of an RFP normalization control *in trans*, analogous to a dual-luciferase assay set-up, revealed high DENR activity (stuORF–GFP/normalization-RFP) in proliferating tissues (brain and imaginal discs), and low activity in tissues with growing, but non-proliferating, cells (salivary gland and fat body, Extended Data Fig. 9).

We wondered how DENR–MCT-1 activity is regulated. Neither DENR protein levels nor DENR–MCT-1 binding dropped in quiescent S2 cells (Extended Data Fig. 10a–c). Phosphorylation of T82, T125 and a double phosphorylation on T118 and S119 in human and fly MCT-1 have been observed[25]. Using cells where endogenous MCT-1 is knocked down via its 3′ UTR and then reconstituted with MCT-1 versions lacking the endogenous 3′ UTR revealed that mutations blocking T118/S119 phosphorylation abolished MCT-1 activity (Extended Data Fig. 10d, d′). Notably, T118 and S119 are evolutionarily conserved in humans. Although MCT-1 was observed to be phosphorylatable *in vitro* by Erk and Cdc2 (ref. 26), we could not observe an effect of Erk, Cdc2, PI(3)K, Akt or TORC1 inhibition on stuORF reporter expression (Extended Data Fig. 10e–g). Further work will be required to identify upstream kinases regulating DENR–MCT-1.

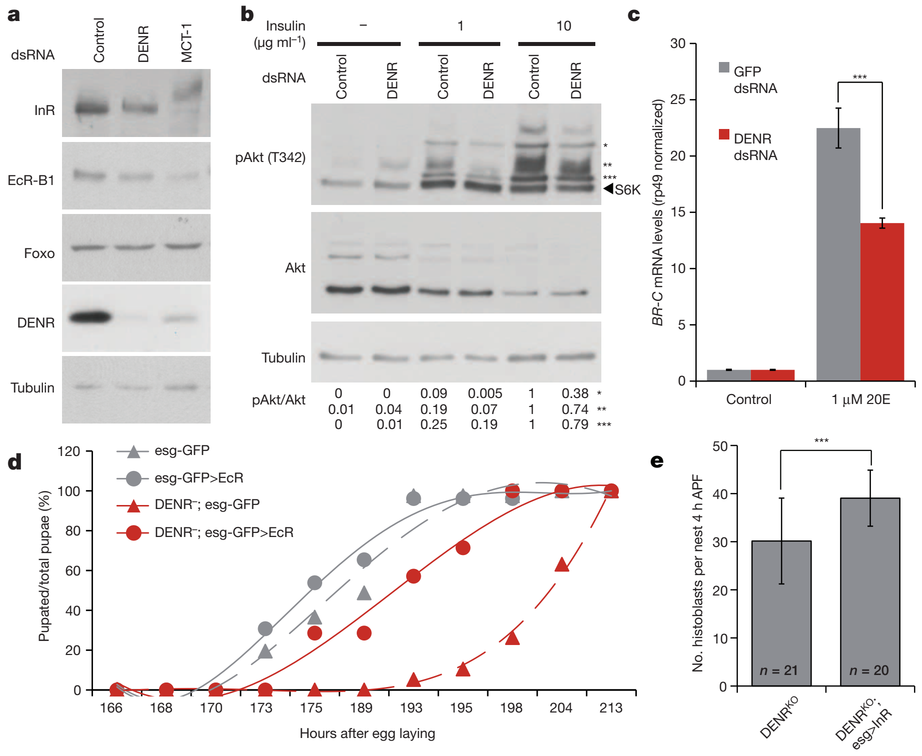


Figure 4 | Loss of DENR leads to reduced InR and EcR protein levels and signalling. a, DENR and MCT-1 knockdown cells have reduced InR and EcR protein levels. **b**, **c**, DENR knockdown cells are less sensitive to insulin stimulation (1 h) (**b**) and to ecdysone (20E, 1 μM, 4 h; **c**). **d**, **e**, Expression of EcR (**d**) or InR (**e**) in histoblast cells and imaginal discs of DENR^{KO} using escargot-GAL4 rescues their delayed pupation (**d**) and mildly but significantly their proliferation defect (**e**). Error bars indicate s.d. *t*-test $***P < 0.001$.

We have identified a new translational control system regulating an abundant class of mRNAs, featuring: (1) stuORFs as the critical *cis*-element; (2) DENR–MCT-1 as the *trans*-acting factor; and (3) proliferation as an important cellular context. This system differs fundamentally from GCN4/ATF4 paradigms both mechanistically and functionally. Unlike GCN4-type mechanisms[1,2,5,27–30], DENR–MCT-1 functions in non-stressed cells, when general translation is not compromised, and independently of uORF to main-ORF distance (Fig. 2h), to promote proliferation. Importantly, DENR–MCT-1 uncouples translation re-initiation from standard initiation, as it is not required for initiation (Fig. 2c (right)). In contrast, GCN4-type mechanisms rely on coupling of initiation and re-initiation to antagonistically regulate GCN4/ATF4 versus all other genes (Supplementary discussion). Our results suggest that re-initiation can be independently controlled via DENR–MCT-1 to modulate translation of a specific group of mRNAs.

METHODS SUMMARY

DENR knockout flies were generated by homologous recombination using pW25. Phospho-dAkt(T342) antibody was developed in collaboration with PhosphoSolutions. Luciferase assays were performed using a dual-luciferase set-up based on pGL3 vectors containing either firefly or *Renilla* luciferase, and the *Drosophila hsp70* basal promoter. The number of replicates for each experiment are described in Supplementary Information.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 2 July 2013; accepted 23 April 2014.
Published online 6 July 2014.

1. Jackson, R. J., Hellen, C. U. & Pestova, T. V. The mechanism of eukaryotic translation initiation and principles of its regulation. *Nature Rev. Mol. Cell Biol.* **11,** 113–127 (2010).
2. Sonenberg, N. & Hinnebusch, A. G. Regulation of translation initiation in eukaryotes: mechanisms and biological targets. *Cell* **136,** 731–745 (2009).
3. Calvo, S. E., Pagliarini, D. J. & Mootha, V. K. Upstream open reading frames cause widespread reduction of protein expression and are polymorphic among humans. *Proc. Natl Acad. Sci. USA* **106,** 7507–7512 (2009).
4. Dever, T. E. & Green, R. The elongation, termination, and recycling phases of translation in eukaryotes. *Cold Spring Harb. Perspect. Biol.* **4,** a013706 (2012).
5. Valasek, L. S. 'Ribozoomin'–translation initiation from the perspective of the ribosome-bound eukaryotic initiation factors (eIFs). *Curr. Protein Pept. Sci.* **13,** 305–330 (2012).
6. Skabkin, M. A., Skabkina, O. V., Hellen, C. U. & Pestova, T. V. Reinitiation and other unconventional posttermination events during eukaryotic translation. *Mol. Cell* **51,** 249–264 (2013).
7. Poyry, T. A., Kaminski, A. & Jackson, R. J. What determines whether mammalian ribosomes resume scanning after translation of a short upstream open reading frame? *Genes Dev.* **18,** 62–75 (2004).
8. Schwanhäusser, B. *et al.* Global quantification of mammalian gene expression control. *Nature* **473,** 337–342 (2011).
9. Gebauer, F. & Hentze, M. W. Molecular mechanisms of translational control. *Nature Rev. Mol. Cell Biol.* **5,** 827–835 (2004).
10. Kong, J. & Lasko, P. Translational control in cellular and developmental processes. *Nature Rev. Genet.* **13,** 383–394 (2012).
11. Araujo, P. R. *et al.* Before it gets started: regulating translation at the 5′ UTR. *Comp. Funct. Genomics* **2012,** 475731 (2012).
12. Ingolia, N. T., Lareau, L. F. & Weissman, J. S. Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. *Cell* **147,** 789–802 (2011).
13. Hinnebusch, A. G. & Lorsch, J. R. The mechanism of eukaryotic translation initiation: new insights and challenges. *Cold Spring Harb. Perspect. Biol.* **4,** 1–25 (2012).
14. Jackson, R. J., Hellen, C. U. & Pestova, T. V. Termination and post-termination events in eukaryotic translation. *Adv. Protein Chem. Struct. Biol.* **86,** 45–93 (2012).
15. Somers, J., Poyry, T. & Willis, A. E. A perspective on mammalian upstream open reading frame function. *Int. J. Biochem. Cell Biol.* **45,** 1690–1700 (2013).
16. Dmitriev, S. E. *et al.* GTP-independent tRNA delivery to the ribosomal P-site by a novel eukaryotic translation factor. *J. Biol. Chem.* **285,** 26779–26787 (2010).
17. Skabkin, M. A. *et al.* Activities of Ligatin and MCT-1/DENR in eukaryotic translation initiation and ribosomal recycling. *Genes Dev.* **24,** 1787–1801 (2010).
18. Dierov, J., Prosniak, M., Gallia, G. & Gartenhaus, R. B. Increased G1 cyclin/cdk activity in cells overexpressing the candidate oncogene, MCT-1. *J. Cell. Biochem.* **74,** 544–550 (1999).
19. Mazan-Mamczarz, K. *et al.* Targeted suppression of MCT-1 attenuates the malignant phenotype through a translational mechanism. *Leuk. Res.* **33,** 474–482 (2009).
20. Prosniak, M. *et al.* A novel candidate oncogene, MCT-1, is involved in cell cycle progression. *Cancer Res.* **58,** 4233–4237 (1998).

21. Kongsuwan, K. *et al.* A *Drosophila Minute* gene encodes a ribosomal protein. *Nature* **317,** 555–558 (1985).
22. Hayashi, S., Hirose, S., Metcalfe, T. & Shirras, A. D. Control of imaginal cell development by the *escargot* gene of *Drosophila. Development* **118,** 105–115 (1993).
23. Wilson, T. G., Yerushalmi, Y., Donnell, D. M. & Restifo, L. L. Interaction between hormonal signaling pathways in *Drosophila melanogaster* as revealed by genetic interaction between methoprene-tolerant and broad-complex. *Genetics* **172,** 253–264 (2006).
24. Huang da, W., Sherman, B. T. & Lempicki, R. A. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.* **37,** 1–13 (2009).
25. Bodenmiller, B. *et al.* PhosphoPep–a phosphoproteome resource for systems biology research in *Drosophila* Kc167 cells. *Mol. Syst. Biol.* **3,** 139 (2007).
26. Nandi, S. *et al.* Phosphorylation of MCT-1 by p44/42 MAPK is required for its stabilization in response to DNA damage. *Oncogene* **26,** 2283–2289 (2007).
27. Spriggs, K. A., Bushell, M. & Willis, A. E. Translational regulation of gene expression during conditions of cell stress. *Mol. Cell* **40,** 228–237 (2010).
28. Harding, H. P. *et al.* Regulated translation initiation controls stress-induced gene expression in mammalian cells. *Mol. Cell* **6,** 1099–1108 (2000).
29. Vattem, K. M. & Wek, R. C. Reinitiation involving upstream ORFs regulates ATF4 mRNA translation in mammalian cells. *Proc. Natl Acad. Sci. USA* **101,** 11269–11274 (2004).
30. Hood, H. M., Neafsey, D. E., Galagan, J. & Sachs, M. S. Evolutionary roles of upstream open reading frames in mediating gene regulation in fungi. *Annu. Rev. Microbiol.* **63,** 385–409 (2009).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank B. Bukau, R. Green and P. Soba for suggestions on the manuscript, V. Benes and T. Bähr-Ivacevic (EMBL Genomics Core Facility) for assistance with microarray experiments, S. Abmayr for anti-Mbc antibody, T. Hsu for anti-Awd antibody, C. Strein and M. Hentze for *Drosophila* DENR antibodies, P. Jakob for help with testing conditions for S2 *in vitro* translation extracts, L. Schibalski and J. Grawe for help with cloning, S. Hofmann for help with polysome analysis, and S. Lerch, A. Haffner and M. Schröder for technical assistance. K.K.M. is the recipient of an Alzheimer's Research Scholarship from the Hans und Ilse Breuer Foundation. P.C.J. is supported in part by a grant from the Fritz Thyssen Foundation to K.E.D. This work was supported in part by a Deutsche Forschungsgemeinschaft (DFG) grant and ERC Starting Grant to A.A.T.

Author Contributions S.S. performed most experiments, except as indicated below. K.S. analysed histoblast proliferation, EcR and InR protein levels and signalling in cells and animals, and performed EcR/InR rescue experiments *in vivo*. P.C.J. assessed proliferation, translation, rRNA and tRNA levels, as well as polysome profiles and genome-wide polysomal profiling of DENR knockdown S2 cells, and performed *in vitro* translation assays. T.K. established and performed inducible reporter assays in proliferating and quiescent cells (Extended Data Figs 7a–f and 10c). K.K.M. performed DENR–MCT-1 co-immunoprecipitation assays (Extended Data Fig. 10b). K.H. performed DENR RNA immunoprecipitation assays (Extended Data Fig. 4g). Y.-S.C. analysed MCT-1 levels and phosphorylation (Extended Data Fig. 10d′). K.K. established *in vitro* translation assays from S2 cells. A.A.T. performed the bioinformatic analyses. K.E.D. and A.A.T. jointly designed and coordinated the study and wrote the paper with input from G.S. All authors interpreted data, discussed results and contributed to writing the manuscript.

Author Information Polysome microarray data are deposited at NCBI GEO under accession number GSE54625. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to K.E.D. (kent.duncan@zmnh.uni-hamburg.de) or A.A.T. (a.teleman@dkfz.de).

doi:10.1038/nature13380

Histone H4 tail mediates allosteric regulation of nucleosome remodelling by linker DNA

William L. Hwang[1,2,3]*, Sebastian Deindl[1,4]*, Bryan T. Harada[1,2] & Xiaowei Zhuang[1,4,5]

Imitation switch (ISWI)-family remodelling enzymes regulate access to genomic DNA by mobilizing nucleosomes[1]. These ATP-dependent chromatin remodellers promote heterochromatin formation and transcriptional silencing[1] by generating regularly spaced nucleosome arrays[2–5]. The nucleosome-spacing activity arises from the dependence of nucleosome translocation on the length of extranucleosomal linker DNA[6–10], but the underlying mechanism remains unclear. Here we study nucleosome remodelling by human ATP-dependent chromatin assembly and remodelling factor (ACF), an ISWI enzyme comprising a catalytic subunit, Snf2h, and an accessory subunit, Acf1 (refs 2, 11–13). We find that ACF senses linker DNA length through an interplay between its accessory and catalytic subunits mediated by the histone H4 tail of the nucleosome. Mutation of AutoN, an auto-inhibitory domain within Snf2h that bears sequence homology to the H4 tail[14], abolishes the linker-length sensitivity in remodelling. Addition of exogenous H4-tail peptide or deletion of the nucleosomal H4 tail also diminishes the linker-length sensitivity. Moreover, Acf1 binds both the H4-tail peptide and DNA in an amino (N)-terminal domain dependent manner, and in the ACF-bound nucleosome, lengthening the linker DNA reduces the Acf1-H4 tail proximity. Deletion of the N-terminal portion of Acf1 (or its homologue in yeast) abolishes linker-length sensitivity in remodelling and leads to severe growth defects *in vivo*. Taken together, our results suggest a mechanism for nucleosome spacing where linker DNA sensing by Acf1 is allosterically transmitted to Snf2h through the H4 tail of the nucleosome. For nucleosomes with short linker DNA, Acf1 preferentially binds to the H4 tail, allowing AutoN to inhibit the ATPase activity of Snf2h. As the linker DNA lengthens, Acf1 shifts its binding preference to the linker DNA, freeing the H4 tail to compete AutoN off the ATPase and thereby activating ACF.

The packaging of DNA into nucleosomes presents a substantial energy barrier that restricts access to the genomic DNA[15]. ISWI-family remodellers use the energy from ATP hydrolysis to disrupt histone–DNA contacts and reposition nucleosomes[1]. The catalytic subunits of ISWI enzymes possess an SF2-like ATPase that translocates DNA across the nucleosome[1]. The nucleosome translocation activity is further regulated by the accessory subunits of ISWI complexes[6,10,16]. Many ISWI remodellers exhibit a nucleosome-spacing activity[2–5]. Critical to this spacing activity are two features of the nucleosome that modulate the activity of ISWI remodellers: (1) the N-terminal tail of histone H4 (refs 8, 17–20) and (2) the length of the extranucleosomal linker DNA[6–10]. The unmodified H4 tail stimulates ISWI activity by relieving the autoinhibitory effect of the AutoN domain within the catalytic subunit[14]. H4 tail acetylation associated with transcriptionally active chromatin is thought to help prevent ISWI-induced nucleosome spacing at actively transcribed genes[17–19]. Regulation by the extranucleosomal linker DNA is responsible for generating the regularly spaced nucleosome arrays important for heterochromatin formation. Shortening the linker DNA reduces the remodelling activity of nucleosome-spacing ISWI enzymes[6–10]. As a result, nucleosomes are preferentially moved towards longer linkers to promote uniform spacing on nucleosome arrays. Interestingly, the catalytic activity of many ISWI-family enzymes is sensitive to linker DNA lengths up to approximately 60–70 base pairs (bp)[6–10], consistent with the inter-nucleosome spacing of heterochromatin observed in human cells[21]. This linker-length sensing range substantially exceeds the binding footprint (20–30 bp) of the catalytic subunit[22,23], whereas the accessory subunits of ISWI complexes can bind linker DNA as far as ~60 bp from the nucleosome edge[22]. However, it is unknown how accessory subunits communicate linker length information to the catalytic subunit to regulate remodelling activity. In this work, we investigate the mechanism underlying DNA linker-length sensing by a prototypical ISWI-family enzyme, human ACF.

To examine how linker DNA regulates nucleosome translocation by ACF, we reconstituted mononucleosomes with varying linker lengths ($n = 20$–78 bp) on the entry side but a constant exit-side linker length of 3 bp (Fig. 1a). We also constructed mononucleosomes with wild-type (WT) histone H4 and two H4 mutants: (1) H4 tail deletion (H4Δ1–19) and (2) H4 with K16A mutation (H4K16A). We refer to nucleosome constructs with the following nomenclature: [WT H4/H4Δ1–19/H4K16A, *n* bp] for nucleosomes with *n* bp of DNA on the entry side and an octamer containing WT H4, H4Δ1–19 or H4K16A. We detected ACF-catalysed nucleosome translocation using fluorescence resonance energy transfer (FRET) by labelling the end of the exit-side linker DNA with the FRET acceptor Cy5, and the histone H2A with the FRET donor Cy3 (Fig. 1a)[24].

We first compared the remodelling kinetics of [WT H4, 78 bp], [WT H4, 40 bp], [WT H4, 20 bp] and [H4Δ1–19, 78 bp] nucleosomes using an ensemble FRET assay[6]. Upon addition of ACF and ATP, the FRET efficiency decreased as DNA was translocated towards the exit side (Fig. 1b and Extended Data Fig. 1a). As expected, the remodelling rate decreased as the linker DNA was shortened and deletion of the H4 tail drastically reduced the remodelling activity (Fig. 1b).

To identify which step(s) of the remodelling process are regulated, we monitored the remodelling of individual nucleosomes using single-molecule FRET[24,25]. Single-nucleosome remodelling traces featured incremental translocation of DNA to the exit side interrupted by kinetic pauses (Fig. 1c). The first pause occurred after ~7 bp of DNA translocation and the second pause occurred after an additional ~3 bp of translocation (Extended Data Fig. 2a, b), consistent with previous findings[24,26]. Moreover, the step sizes did not change with linker DNA length or histone H4 modification (Extended Data Fig. 2a, b). We divided the remodelling time trace into two translocation phases (T1, T2), during which the FRET efficiency decreased, and two pause phases (P1, P2), during which the FRET value remained constant (Fig. 1c). Notably, the DNA translocation rates between pauses did not change, whereas the pause-phase exit rates decreased dramatically when the linker DNA was shortened (Fig. 1d and Extended Data Fig. 2c). Moreover, the dependence of remodelling kinetics on entry-side linker lengths of mononucleosomes was quantitatively similar to the dependence on inter-nucleosome linker

[1] Howard Hughes Medical Institute, Harvard University, Cambridge, Massachusetts 02138, USA. [2] Graduate Program in Biophysics, Harvard University, Cambridge, Massachusetts 02138, USA. [3] Harvard/MIT MD-PhD Program, Harvard Medical School, Boston, Massachusetts 02115, USA. [4] Department of Chemistry and Chemical Biology, Harvard University, Cambridge, Massachusetts 02138, USA. [5] Department of Physics, Harvard University, Cambridge, Massachusetts 02138, USA.

*These authors contributed equally to this work.

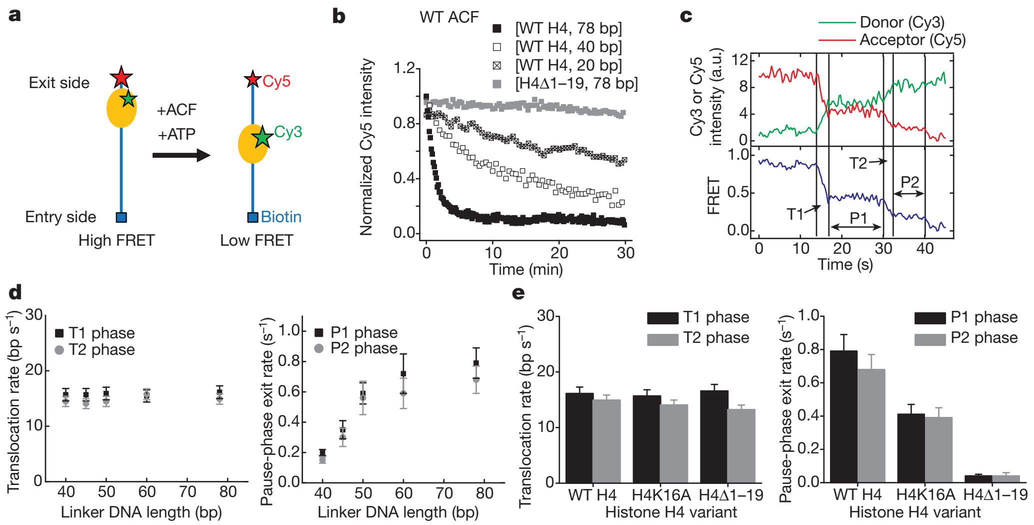


Figure 1 | The linker DNA length and histone H4 tail regulate the remodelling pause phases but not the translocation phases. a, Schematic of a FRET-labelled mononucleosome undergoing remodelling by ACF. **b**, Ensemble remodelling time courses of [WT H4, 78 bp], [WT H4, 40 bp], [WT H4, 20 bp] and [H4Δ1–19, 78 bp] nucleosomes by 40 nM ACF at 5 μM ATP. Nucleosome translocation is monitored by the emission intensity of the FRET acceptor Cy5 under excitation of the FRET donor Cy3. **c**, Cy3 and Cy5 fluorescence (top; a.u., arbitrary units) and FRET (bottom) time traces during the remodelling of a single [WT H4, 78 bp] nucleosome with the translocation (T1, T2) and pause (P1, P2) phases indicated. **d**, Linker DNA length dependence of the translocation rates between pauses (left, defined as the average number of base pairs moved per second) and pause-phase exit rates (right, defined as the inverse of the average pause durations). **e**, Dependence of the translocation rates between pauses (left) and pause-phase exit rates (right) on the H4 variants. In **d** and **e**, [ACF] = 10 nM and [ATP] = 2 mM. Data are mean ± s.e.m. derived from at least 100 (**d**) or at least 50 (**e**) individual nucleosome remodelling traces from three independent experiments.

lengths of dinucleosomes (Extended Data Fig. 3), validating the use of mononucleosomes as a model system to study linker-length sensitivity. Interestingly, the H4 tail appeared to regulate the same phase of the remodelling process as the linker DNA (Fig. 1e). The H4K16A mutation and H4 tail deletion (H4Δ1–19) decreased the pause-phase exit rate by approximately 2- and 20-fold, respectively (Fig. 1e). In contrast, neither modification had any appreciable effect on the translocation rates between pauses (Fig. 1e).

The above results indicate that both linker DNA and the H4 tail regulate the remodelling rate by changing the duration of pause phases, suggesting that these nucleosome features may impinge on an inhibitory mechanism that prevents the initiation of the DNA translocation phases. It has been shown that although the ISWI ATPase domain can translocate nucleosomes autonomously[27], the catalytic subunit contains two well-conserved autoregulatory domains, AutoN and NegC, which inhibit ATP hydrolysis and its coupling to DNA translocation, respectively[14]. The AutoN inhibition can be relieved by the H4 tail whereas the NegC inhibition can be relieved by binding of the HAND-SANT-SLIDE module to linker DNA[14]. Could the regulation of remodelling by linker DNA length occur through these inhibitory domains?

To address this question, we first examined the role of the NegC domain. Surprisingly, deletion of the NegC domain in the ACF complex (ΔNegC ACF) did not substantially affect the dependence of remodelling kinetics on linker DNA lengths ranging from 20 to 78 bp (Fig. 2a–c and Extended Data Fig. 4a). Removing the H4 tail dramatically reduced the remodelling rates of both WT and ΔNegC ACF (Fig. 2b). These results suggest that the NegC domain does not play a substantial role in linker length sensing by the ACF complex. In contrast, the isolated Snf2h catalytic subunit exhibited a short-range (20–40 bp) linker length sensitivity that depended on NegC (Extended Data Fig. 5), in a manner similar to the *Drosophila* ISWI, which lacks any accessory subunit[14].

Next, we mutated the AutoN domain with two point substitutions (R142A and R144A) in the ACF complex (AutoN-2RA ACF; Fig. 3a and Extended Data Fig. 4a). AutoN bears sequence homology to the H4 tail, which can compete the inhibitory AutoN domain off the ATPase, and the 2RA mutation in AutoN is expected to diminish the H4 tail dependence of remodelling by ISWI enzymes[14]. Remarkably, this mutation not only increased the remodelling rate of nucleosomes lacking the H4 tail, but also completely abolished the linker-length dependence of remodelling by specifically increasing the remodelling rate of short-linker nucleosomes (Fig. 3b, c and Extended Data Fig. 6). These results suggest an essential role for AutoN in linker length sensing by the ACF complex.

Since AutoN competes with the H4 tail for binding to the ATPase[14], we considered the possibility that this competition is involved in sensing linker DNA length and hypothesized that the H4 tail is only available to compete AutoN off the ATPase when the linker DNA is sufficiently

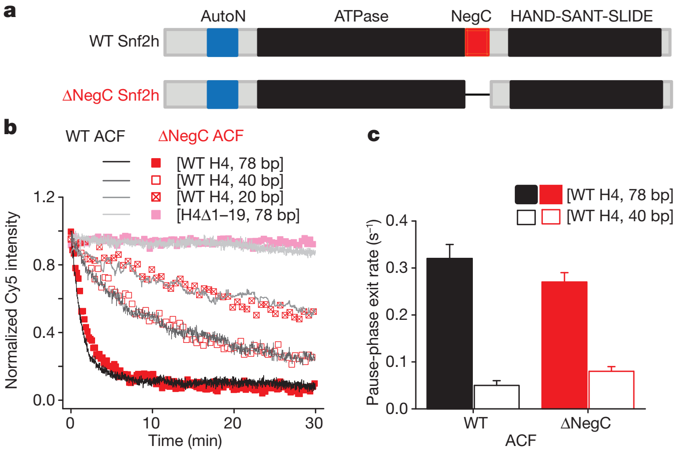


Figure 2 | Deletion of the NegC domain of the Snf2h catalytic subunit does not substantially affect linker DNA length sensing by the ACF complex. a, Domain architecture of WT and ΔNegC Snf2h (residues 669–700 replaced with a SGSGS linker). **b**, Ensemble remodelling time courses of [WT H4, 78 bp], [WT H4, 40 bp], [WT H4, 20 bp] and [H4Δ1–19, 78 bp] nucleosomes by 40 nM WT ACF (black/grey lines, duplicated from Fig. 1b) and ΔNegC ACF (red/pink symbols) at 5 μM ATP. **c**, Linker DNA length dependence of the pause-phase exit rate (P1 phase) measured for WT ACF (black) and ΔNegC ACF (red). [ACF] = 10 nM and [ATP] = 20 μM. Data are mean ± s.e.m. derived from at least 100 individual nucleosome remodelling traces from three independent experiments.

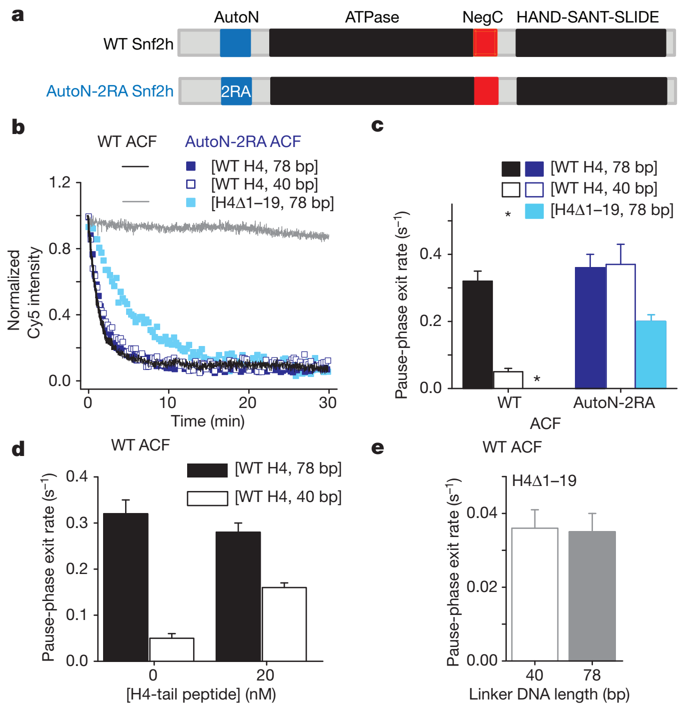


Figure 3 | The AutoN domain of Snf2h and the nucleosomal H4 tail are important for linker DNA length sensing by the ACF complex. **a**, Domain architecture of WT and AutoN-2RA (R142A and R144A) Snf2h. **b**, Ensemble remodelling time courses of [WT H4, 78 bp], [WT H4, 40 bp] and [H4Δ1–19, 78 bp] nucleosomes by 40 nM WT ACF (black/grey lines, duplicated from Fig. 1b) and AutoN-2RA ACF (blue/cyan symbols) at 5 μM ATP. **c**, Dependence of the pause-phase exit rate on the linker DNA length and H4 tail for WT (black) or AutoN-2RA ACF (blue/cyan). *Too slow to be measured. **d**, Effect of the exogenously added H4-tail peptide on the pause-phase exit rates during remodelling by WT ACF. **e**, Pause-phase exit rates of nucleosomes lacking the H4 tail during remodelling by WT ACF. In **c**–**f**, [ACF] = 10 nM and [ATP] = 20 μM, except that 2 mM of ATP was used in **e** to make the pause exit rates measurable for nucleosomes lacking the H4 tail. Data are mean ± s.e.m. from at least 100 (**c**, **d**) or at least 50 (**e**) individual nucleosome remodelling traces from three independent experiments.

long. Consistent with this hypothesis, adding exogenous H4-tail peptide, which should help compete AutoN off the ATPase when the nucleosomal H4 tail is unavailable, specifically increased the remodelling rate of short-linker nucleosome ([WT H4, 40 bp]) by WT ACF (Fig. 3d). Furthermore, deletion of the nucleosomal H4 tail, in addition to slowing down remodelling, abolished the dependence of remodelling rate on linker DNA length (Fig. 3e). These results indicate that the H4 tail is indeed involved in linker DNA sensing.

Because the catalytic subunits of ISWI-family enzymes only interact with ~20–30 bp of extranucleosomal DNA, the linker-length sensitivity of ACF cannot be accounted for by the catalytic subunit alone. Our findings raise the intriguing possibility of a linker-length sensing mechanism where the accessory subunit Acf1 interacts with the H4 tail in a linker-length-dependent manner, which modulates the H4 tail availability for competing with AutoN. To test this possibility, we generated two Acf1 mutants, ΔC-term Acf1 and ΔN-term Acf1, in which 134 residues at the carboxy (C) terminus or 371 residues at the N terminus were deleted, respectively (Extended Data Fig. 7a and Fig. 4a). Because the central region of Acf1 required for Snf2h binding[16,28] was not deleted, both mutants were able to form complexes with Snf2h, which are referred to as ΔC-term and ΔN-term ACF (Extended Data Fig. 4b).

We first probed which region of Acf1 interacts with the H4 tail by comparing the binding affinities of WT, ΔC-term and ΔN-term Acf1 for the H4-tail peptide using a fluorescence anisotropy assay. Interestingly, WT Acf1 exhibited specific, nanomolar affinity for the H4-tail peptide (Fig. 4b and Extended Data Fig. 7b) that was not substantially altered upon deletion of the C-terminal region (Extended Data Fig. 7c), but was completely lost upon deletion of the N-terminal portion (Fig. 4b). These results indicate that Acf1 interacts with the H4 tail probably through its N-terminal region. Acf1 also bound double-stranded DNA and deletion of the N-terminal region abolished this interaction too (Extended Data Fig. 8), consistent with the previous finding that the WAC motif within the N-terminal region is important for binding of ACF to the linker DNA[28]. Given the distinct properties of DNA and the H4 tail, their specific binding interfaces within Acf1 N-term are probably distinct.

Next, we investigated nucleosome remodelling by the ΔC-term and ΔN-term ACF complexes. Notably, the ΔC-term mutation did not substantially alter the dependence of remodelling kinetics on linker DNA length (Extended Data Fig. 7d), whereas the linker-length sensitivity was eliminated in the ΔN-term ACF complex (Fig. 4c, d). This finding is consistent with the specific affinity of Acf1 N-term for the H4 tail (Fig. 4b). Furthermore, if the loss of linker-length sensitivity was simply a result of losing the linker DNA binding affinity of Acf1, ΔN-term ACF should demonstrate inefficient remodelling for all linker DNA lengths. Instead, ΔN-term ACF remodelled both short- and long-linker nucleosomes at rates close to the rate with which WT ACF remodelled long-linker nucleosomes (Fig. 4c, d), suggesting that deletion of Acf1 N-term disabled a mechanism that inhibits remodelling at short linker lengths. ΔN-term ACF also maintained the H4-tail requirement in remodelling (Fig. 4c).

Since Acf1 has affinity to both DNA and the H4 tail, a plausible interpretation of the above observations is that the nucleosomal linker DNA and H4 tail are in competition for binding to the N-terminal region of Acf1 and that this competition is modulated by the length of the linker DNA. Only when the linker is sufficiently short does Acf1 preferentially bind to the H4 tail, making it unavailable to compete with the inhibitory AutoN. Deletion of Acf1 N-term diminishes the Acf1-H4 tail interaction such that the H4 tail is equally available to activate the ATPase at both short and long linker DNA lengths. We therefore probed the linker-length dependence of the Acf1-H4 tail proximity in ACF-bound nucleosomes featuring a cysteine-reactive crosslinker on the H4 tail. Specific H4-Acf1 crosslinking product was clearly observed as a band with reduced electrophoretic mobility compared with non-crosslinked Acf1 (Fig. 4e and Extended Data Fig. 9). Remarkably, the Acf1-H4 crosslinking efficiency decreased substantially with increasing linker DNA length (Fig. 4e), supporting our hypothesis that the Acf1-H4 tail interaction is modulated by the linker DNA length. In contrast, the H4-Snf2h crosslinking efficiency did not change substantially with linker DNA length, probably because Snf2h remains sufficiently close to the H4 tail regardless of the linker DNA length, which allows crosslinking even when the H4 tail was not specifically bound to its binding pocket on Snf2h.

Finally, we tested the physiological importance of the N-terminal region of Acf1 by studying the role of its homologue in yeast[29]. Yeast ISW2 is functionally similar to ACF. It is composed of a catalytic subunit (Isw2) that is homologous to Snf2h and three accessory subunits (Itc1, Dpb4 and Dls1), among which Itc1 is homologous to Acf1. We generated three mutant yeast strains: (1) deletion of the entire *itc1* gene (Δ*itc1*), (2) deletion of only the portion of *itc1* that encodes the N-terminal region of Itc1 equivalent to Acf1 N-term (Δ*itc1-Nterm*) and (3) a rescue strain that was derived from the Δ*itc1-Nterm* strain by deleting the remainder of *itc1* (rescue-Δ*itc1*). Both Δ*itc1* and rescue-Δ*itc1* showed growth rates similar to that of the WT strain (Fig. 4f), consistent with previous observations[29]. In contrast, the Δ*itc1-Nterm* strain displayed dramatically slower growth (Fig. 4f), consistent with an aberrant chromatin-misregulation phenotype.

Taken together, our results suggest a nucleosome-spacing mechanism for ACF in which the linker DNA length is sensed by the Acf1 accessory subunit and allosterically transmitted to the Snf2h catalytic subunit through the H4 tail of the nucleosome (Fig. 4g). Acf1 and the AutoN domain of Snf2h function collectively in DNA linker-length sensing. When the linker DNA is short, Acf1 preferentially binds to and sequesters the H4 tail, making it unavailable to compete its sequence homologue, AutoN, off the ATPase. Hence, the ATPase activity is inhibited by AutoN. As

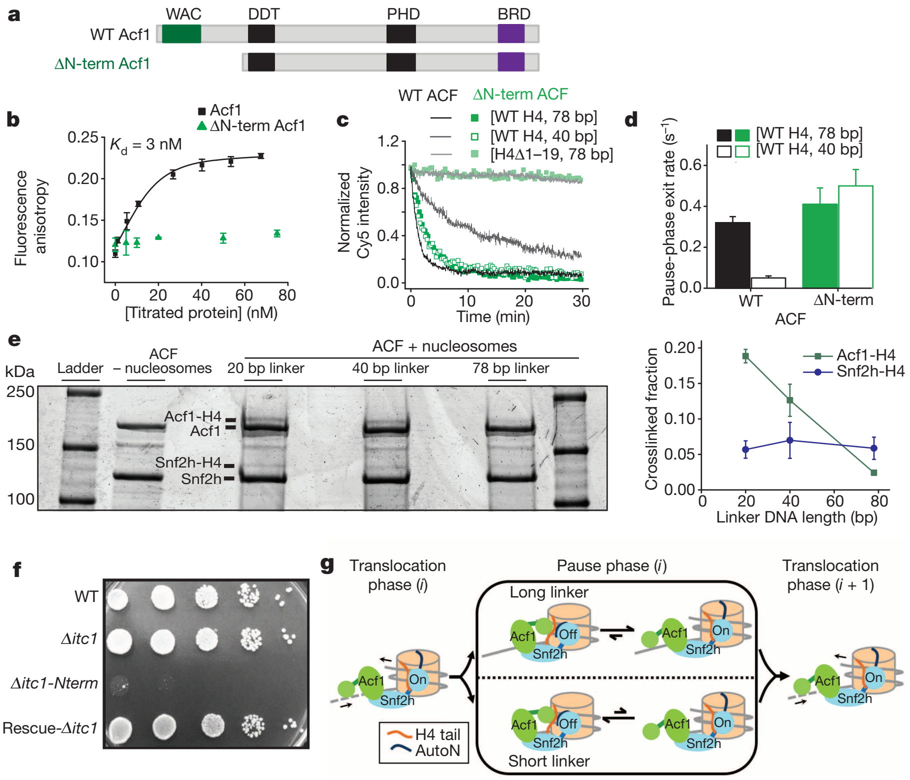


Figure 4 | The N-terminal region of the Acf1 accessory subunit is important for linker DNA length sensing by the ACF complex. a, Domain architecture of WT and ΔN-term (residues 1–371 deleted) Acf1. **b**, Fluorescence anisotropy of dye-labelled H4-tail peptide in the presence of varying amounts of WT (black symbols) or ΔN-term (green symbols) Acf1. Data are mean ± s.e.m. ($n = 3$ independent experiments). The dissociation constant (K_d) for WT Acf1 is 3 ± 9 nM (error bars, 95% confidence intervals). **c**, Ensemble remodelling time courses of [WT H4, 78 bp], [WT H4, 40 bp] and [H4Δ1–19, 78 bp] nucleosomes by 40 nM WT ACF (black/grey lines, duplicated from Fig. 1b) and ΔN-term ACF (green/light green symbols) at 5 μM ATP. **d**, Dependence of the pause-phase exit rate on the linker DNA length for WT ACF (black) or ΔN-term ACF (green). [ACF] = 10 nM and [ATP] = 20 μM. Data are mean ± s.e.m. derived from at least 100 individual nucleosome remodelling traces from three independent experiments. **e**, Crosslinking of the H4 tail to Acf1 depends on the linker DNA length. The crosslinking products were analysed by SDS–PAGE (left). The Acf1-H4 crosslinking band was absent for ACF without nucleosomes (lane 'ACF−nucleosomes'). Right: quantification of the H4-crosslinked fractions of Acf1 and Snf2h as a function of linker DNA length. Data are mean ± s.e.m. ($n = 3$ independent crosslinking experiments). **f**, Effects of deletion of Itc1 (Acf1 homologue) and its N-terminal region on the growth of yeast cells. Top row: WT. Second row: the *itc1* gene is deleted (Δ*itc1*). Third row: the coding sequence of the N-terminal region of Itc1 is deleted (Δ*itc1-Nterm*). Bottom row: the remaining portion of *itc1* is deleted from Δ*itc1-Nterm* (rescue-Δ*itc1*). One representative of three independent growth experiments is shown. **g**, Model for linker DNA length sensing by the ACF complex. DNA: grey lines; histone octamer: beige cylinders; Snf2h: blue/cyan; Acf1: green. The ATPase domain of Snf2h is depicted as a cyan sphere and labelled 'On' when active and 'Off' when inactive.

the linker DNA length increases, Acf1 shifts its binding preference to the linker DNA and releases the H4 tail, allowing it to compete AutoN off the ATPase and activate ACF. This competition between the H4 tail and linker DNA for Acf1 binding probably involves the N-terminal region of Acf1. It is interesting to note that linker DNA sensing occurs during the pause phases when the ATPase domain is not actively translocating DNA, suggesting that AutoN engages the ATPase domain during the pauses. To exit the pauses, the H4 tail is required to relieve the inhibitory effect of AutoN. The re-engagement of AutoN with the ATPase domain after each translocation phase would give ACF an opportunity to periodically sense the linker DNA length. Such frequent sensing may allow a more efficient nucleosome spacing, as previously hypothesized[30]. The linker DNA and the H4 tail are two important substrate features that regulate nucleosome remodelling by ISWI-family enzymes, the former enabling uniform nucleosome spacing for heterochromatin formation and the latter specifying regions of chromatin for silencing. Our results now reveal an unexpected convergence of the regulatory pathways defined by these two distinct nucleosome features.

METHODS SUMMARY

Detailed descriptions of nucleosome and ACF preparation, as well as single-molecule and ensemble FRET, fluorescence anisotropy, protein crosslinking and yeast experiments, are described in Methods. Briefly, various nucleosome constructs were reconstituted using Cy3-labelled histone octamers and Cy5-labelled DNA with a biotin moiety for surface anchoring. DNA was generated by PCR or by annealing and ligating a set of overlapping, complementary oligonucleotides (Extended Data Fig. 10). Histone octamer, nucleosomes, Acf1, Snf2h and ACF complexes were reconstituted and purified as described previously[6,7,24]. Mutant yeast strains were generated in the BY4741 background. Single-molecule FRET measurements were performed with a custom-built microscope setup. Ensemble FRET used a Cary Eclipse fluorescence spectrophotometer. Fluorescence anisotropy measurements used a SpectraMax microplate reader. Crosslinking experiments used nucleosomes with a cysteine-reactive crosslinker BM(PEG)$_3$ at the H4 tail N terminus and reaction products were analysed by SDS–polyacrylamide gel electrophoresis (SDS–PAGE).

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 9 August 2013; accepted 14 April 2014.

Published online 29 June 2014.

1. Clapier, C. R. & Cairns, B. R. The biology of chromatin remodeling complexes. *Annu. Rev. Biochem.* **78,** 273–304 (2009).
2. Ito, T., Bulger, M., Pazin, M. J., Kobayashi, R. & Kadonaga, J. T. ACF, an ISWI-containing and ATP-utilizing chromatin assembly and remodeling factor. *Cell* **90,** 145–155 (1997).
3. Varga-Weisz, P. D. *et al.* Chromatin-remodelling factor CHRAC contains the ATPases ISWI and topoisomerase II. *Nature* **388,** 598–602 (1997).
4. Tsukiyama, T., Palmer, J., Landel, C. C., Shiloach, J. & Wu, C. Characterization of the imitation switch subfamily of ATP-dependent chromatin-remodeling factors in *Saccharomyces cerevisiae. Genes Dev.* **13,** 686–697 (1999).

5. Langst, G., Bonte, E. J., Corona, D. F. & Becker, P. B. Nucleosome movement by CHRAC and ISWI without disruption or trans-displacement of the histone octamer. *Cell* **97,** 843–852 (1999).
6. Yang, J. G., Madrid, T. S., Sevastopoulos, E. & Narlikar, G. J. The chromatin-remodeling enzyme ACF is an ATP-dependent DNA length sensor that regulates nucleosome spacing. *Nature Struct. Mol. Biol.* **13,** 1078–1083 (2006).
7. He, X., Fan, H. Y., Narlikar, G. J. & Kingston, R. E. Human ACF1 alters the remodeling strategy of SNF2h. *J. Biol. Chem.* **281,** 28636–28647 (2006).
8. Dang, W., Kagalwala, M. N. & Bartholomew, B. Regulation of ISW2 by concerted action of histone H4 tail and extranucleosomal DNA. *Mol. Cell. Biol.* **26,** 7388–7396 (2006).
9. Stockdale, C., Flaus, A., Ferreira, H. & Owen-Hughes, T. Analysis of nucleosome repositioning by yeast ISWI and Chd1 chromatin remodeling complexes. *J. Biol. Chem.* **281,** 16279–16288 (2006).
10. He, X., Fan, H. Y., Garlick, J. D. & Kingston, R. E. Diverse regulation of SNF2h chromatin remodeling by noncatalytic subunits. *Biochemistry* **47,** 7025–7033 (2008).
11. Ito, T. *et al.* ACF consists of two subunits, Acf1 and ISWI, that function cooperatively in the ATP-dependent catalysis of chromatin assembly. *Genes Dev.* **13,** 1529–1539 (1999).
12. LeRoy, G., Loyola, A., Lane, W. S. & Reinberg, D. Purification and characterization of a human factor that assembles and remodels chromatin. *J. Biol. Chem.* **275,** 14787–14790 (2000).
13. Poot, R. A. *et al.* HuCHRAC, a human ISWI chromatin remodelling complex contains hACF1 and two novel histone-fold proteins. *EMBO J.* **19,** 3377–3387 (2000).
14. Clapier, C. R. & Cairns, B. R. Regulation of ISWI involves inhibitory modules antagonized by nucleosomal epitopes. *Nature* **492,** 280–284 (2012).
15. Killian, J. L., Li, M., Sheinin, M. Y. & Wang, M. D. Recent advances in single molecule studies of nucleosomes. *Curr. Opin. Struct. Biol.* **22,** 80–87 (2012).
16. Eberharter, A., Vetter, I., Ferreira, R. & Becker, P. B. ACF1 improves the effectiveness of nucleosome mobilization by ISWI through PHD-histone contacts. *EMBO J.* **23,** 4029–4039 (2004).
17. Clapier, C. R., Langst, G., Corona, D. F., Becker, P. B. & Nightingale, K. P. Critical role for the histone H4 N terminus in nucleosome remodeling by ISWI. *Mol. Cell. Biol.* **21,** 875–883 (2001).
18. Hamiche, A., Kang, J. G., Dennis, C., Xiao, H. & Wu, C. Histone tails modulate nucleosome mobility and regulate ATP-dependent nucleosome sliding by NURF. *Proc. Natl Acad. Sci. USA* **98,** 14316–14321 (2001).
19. Shogren-Knaak, M. *et al.* Histone H4–K16 acetylation controls chromatin structure and protein interactions. *Science* **311,** 844–847 (2006).
20. Ferreira, H., Flaus, A. & Owen-Hughes, T. Histone modifications influence the action of Snf2 family remodelling enzymes by different mechanisms. *J. Mol. Biol.* **374,** 563–579 (2007).
21. Valouev, A. *et al.* Determinants of nucleosome organization in primary human cells. *Nature* **474,** 516–520 (2011).
22. Kagalwala, M. N., Glaus, B. J., Dang, W., Zofall, M. & Bartholomew, B. Topography of the ISW2-nucleosome complex: insights into nucleosome spacing and chromatin remodeling. *EMBO J.* **23,** 2092–2104 (2004).
23. Yamada, K. *et al.* Structure and mechanism of the chromatin remodelling factor ISW1a. *Nature* **472,** 448–453 (2011).
24. Blosser, T. R., Yang, J. G., Stone, M. D., Narlikar, G. J. & Zhuang, X. Dynamics of nucleosome remodelling by individual ACF complexes. *Nature* **462,** 1022–1027 (2009).
25. Ha, T. *et al.* Probing the interaction between two single molecules: fluorescence resonance energy transfer between a single donor and a single acceptor. *Proc. Natl Acad. Sci. USA* **93,** 6264–6268 (1996).
26. Deindl, S. *et al.* ISWI remodelers slide nucleosomes with coordinated multi-base-pair entry steps and single-base-pair exit steps. *Cell* **152,** 442–452 (2013).
27. Mueller-Planitz, F., Klinker, H., Ludwigsen, J. & Becker, P. B. The ATPase domain of ISWI is an autonomous nucleosome remodeling machine. *Nature Struct. Mol. Biol.* **20,** 82–89 (2013).
28. Fyodorov, D. V. & Kadonaga, J. T. Binding of Acf1 to DNA involves a WAC motif and is important for ACF-mediated chromatin assembly. *Mol. Cell. Biol.* **22,** 6344–6353 (2002).
29. Gelbart, M. E., Rechsteiner, T., Richmond, T. J. & Tsukiyama, T. Interactions of Isw2 chromatin remodeling complex with nucleosomal arrays: analyses using recombinant yeast histones and immobilized templates. *Mol. Cell. Biol.* **21,** 2098–2106 (2001).
30. Narlikar, G. J. A proposal for kinetic proof reading by ISWI family chromatin remodeling motors. *Curr. Opin. Chem. Biol.* **14,** 660–665 (2010).

Acknowledgements We thank G. Narlikar, T. Tsukiyama, J. Vaughan, J. Moffitt and E. Sun for discussions, S. Mukherji for assistance in designing the yeast experiments, S. Trauger, K. L. Hwang, R. Magnusson, G. Hilinski and J. McGee for assistance with protein characterization and biochemistry, and G. Narlikar for providing Snf2h and Acf1 expression vectors and purification protocols. This work was supported in part by the National Institutes of Health (GM105637 to X.Z.). W.L.H. acknowledges support from the National Institutes of Health T32G007753 Training Grant. S.D. was a Merck Fellow of the Jane Coffin Childs Foundation. X.Z. is a Howard Hughes Medical Investigator.

Author Contributions W.L.H., S.D. and X.Z. designed the experiments, with input from B.T.H. W.L.H. and S.D. performed the experiments and data analysis. B.T.H. helped prepare the histones and nucleosomes. S.D., W.L.H. and X.Z. wrote the paper, with input from B.T.H. X.Z. oversaw the project.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to X.Z. (zhuang@chemistry.harvard.edu).

doi:10.1038/nature13430

Visualization of arrestin recruitment by a G-protein-coupled receptor

Arun K. Shukla[1]†*, Gerwin H. Westfield[2]*, Kunhong Xiao[1]*, Rosana I. Reis[1], Li-Yin Huang[1], Prachi Tripathi-Shukla[1], Jiang Qian[1], Sheng Li[3], Adi Blanc[1], Austin N. Oleskie[2], Anne M. Dosey[2], Min Su[2], Cui-Rong Liang[4], Ling-Ling Gu[4], Jin-Ming Shan[4], Xin Chen[4], Rachel Hanna[5], Minjung Choi[6], Xiao Jie Yao[1], Bjoern U. Klink[1], Alem W. Kahsai[1], Sachdev S. Sidhu[5], Shohei Koide[7], Pawel A. Penczek[8], Anthony A. Kossiakoff[7], Virgil L. Woods Jr[3]‡, Brian K. Kobilka[9], Georgios Skiniotis[2] & Robert J. Lefkowitz[1,6,10]

G-protein-coupled receptors (GPCRs) are critically regulated by β-arrestins, which not only desensitize G-protein signalling but also initiate a G-protein-independent wave of signalling[1–5]. A recent surge of structural data on a number of GPCRs, including the β_2 adrenergic receptor (β_2AR)–G-protein complex, has provided novel insights into the structural basis of receptor activation[6–11]. However, complementary information has been lacking on the recruitment of β-arrestins to activated GPCRs, primarily owing to challenges in obtaining stable receptor–β-arrestin complexes for structural studies. Here we devised a strategy for forming and purifying a functional human β_2AR–β-arrestin-1 complex that allowed us to visualize its architecture by single-particle negative-stain electron microscopy and to characterize the interactions between β_2AR and β-arrestin 1 using hydrogen–deuterium exchange mass spectrometry (HDX-MS) and chemical crosslinking. Electron microscopy two-dimensional averages and three-dimensional reconstructions reveal bimodal binding of β-arrestin 1 to the β_2AR, involving two separate sets of interactions, one with the phosphorylated carboxy terminus of the receptor and the other with its seven-transmembrane core. Areas of reduced HDX together with identification of crosslinked residues suggest engagement of the finger loop of β-arrestin 1 with the seven-transmembrane core of the receptor. In contrast, focal areas of raised HDX levels indicate regions of increased dynamics in both the N and C domains of β-arrestin 1 when coupled to the β_2AR. A molecular model of the β_2AR–β-arrestin signalling complex was made by docking activated β-arrestin 1 and β_2AR crystal structures into the electron microscopy map densities with constraints provided by HDX-MS and crosslinking, allowing us to obtain valuable insights into the overall architecture of a receptor–arrestin complex. The dynamic and structural information presented here provides a framework for better understanding the basis of GPCR regulation by arrestins.

To facilitate the isolation of a stable β_2AR–β-arrestin complex, we used a modified β_2AR construct with its C terminus replaced by that of the arginine vasopressin type 2 receptor ($AVPR_2$). This chimaeric receptor (β_2V_2R) maintains pharmacological properties identical to the β_2AR, but it binds β-arrestins with higher affinity compared to wild-type β_2AR[12]. We co-expressed β_2V_2R, β-arrestin 1 (1–393) and GRK2^{CAAX} (GRK2 with a membrane-tethering prenylation signal) in insect cells followed by agonist stimulation and affinity purification through the Flag-tagged receptor (Fig. 1a) However, since the isolation of a stable complex was still not feasible (Fig. 1b, lanes 1 and 2), we explored enhancing its stability by adding Fab30, an antibody fragment we previously reported that selectively recognizes and stabilizes the active conformation of β-arrestin 1 (ref. 13). Indeed, incubation of Fab30 with preformed complex in the membrane resulted in a robust purification of the β_2V_2R–β-arrestin-1 complex (Fig. 1b, lanes 5 and 6), whereas a nonspecific Fab (referred to as Fab1) did not support complex stabilization (Fig. 1b, lanes 3 and 4). Complex isolation was only possible in response to an agonist (BI-167107) and not an inverse agonist (ICI-118551) (Fig. 1b, lanes 5 and 6). Furthermore, the efficiency of complex purification using this approach directly mirrors the pharmacological efficacy of the ligand used to stimulate the cells (Fig. 1c). While stimulation of cells with inverse agonists does not yield detectable co-purification of β-arrestin 1, agonists robustly stabilize the complex and partial agonists yield co-purification of β-arrestin 1 at moderate levels. Moreover, the efficiency of complex formation also corresponds to the ligand occupancy of the receptor as reflected by the increasing amount of β-arrestin 1 co-purification with increasing agonist concentrations (Extended Data Fig. 1a, b). The direct correlation of ligand efficacy and occupancy with purification efficiency reflects the fact that this approach yields a complex that depends on both activated receptor conformation and receptor phosphorylation. The purified β_2V_2R–β-arrestin-1–Fab30 complex also exhibited a robust interaction with the purified clathrin terminal domain compared to β-arrestin 1 alone, suggesting that β-arrestin 1 in this complex is in a physiologically relevant and functional conformation (Extended Data Fig. 2)[14–16]. Importantly, this strategy allowed preparative scale purification of a highly stable β_2V_2R–β-arrestin-1–Fab30 complex as assessed by analytical size exclusion chromatography (Fig. 1a, bottom right, green trace, and Extended Data Fig. 1c). In addition to the Fab30-stabilized β_2V_2R–β-arrestin-1 complex, we were also able to obtain equally stable β_2V_2R–β-arrestin-1 complexes using the single-chain variable fragment of Fab30 (ScFv30) (Fig. 1a, bottom right, blue trace).

The interaction of β-arrestins with activated GPCRs is proposed to involve two sequential steps[17]. First, the phosphorylated C terminus of activated GPCRs is thought to engage the N domain of β-arrestins, a high-affinity charge–charge interaction primarily mediated between the phosphates on the receptor tail and basic residues on β-arrestins[13,17]. This first engagement is hypothesized to facilitate activating conformational changes in β-arrestin, leading in turn to additional interactions with the transmembrane core of the receptor[17]. To obtain dynamic structural information on the receptor–β-arrestin complex, we carried out HDX-MS analysis on the purified assembly[18,19]. In addition to the β_2V_2R–β-arrestin-1–Fab30 complex, we used the $AVPR_2$ C-terminal

[1] Department of Medicine, Duke University Medical Center, Durham, North Carolina 27710, USA. [2] Life Sciences Institute and Department of Biological Chemistry, University of Michigan Medical School, Ann Arbor, Michigan 48109, USA. [3] Department of Chemistry, University of California at San Diego, La Jolla, California 92093, USA. [4] School of Pharmaceutical & Life Sciences, Changzhou University, Changzhou, Jiangsu 213164, China. [5] Terrence Donnelly Centre for Cellular and Biomolecular Research, University of Toronto, Toronto, Ontario M5S 3E1, Canada. [6] Department of Biochemistry, Duke University Medical Center, Durham, North Carolina 27710, USA. [7] Department of Biochemistry and Molecular Biology, University of Chicago, Chicago, Illinois 60637, USA. [8] Department of Biochemistry and Molecular Biology, The University of Texas Medical School at Houston, Houston, Texas 77054, USA. [9] Department of Molecular and Cellular Physiology, Stanford University School of Medicine, 279 Campus Drive, Stanford, California 94305, USA. [10] Howard Hughes Medical Institute, Duke University Medical Center, Durham, North Carolina 27710, USA. †Present address: Department of Biological Sciences and Bioengineering, Indian Institute of Technology, Kanpur 208016, India.
*These authors contributed equally to this work.
‡Deceased.

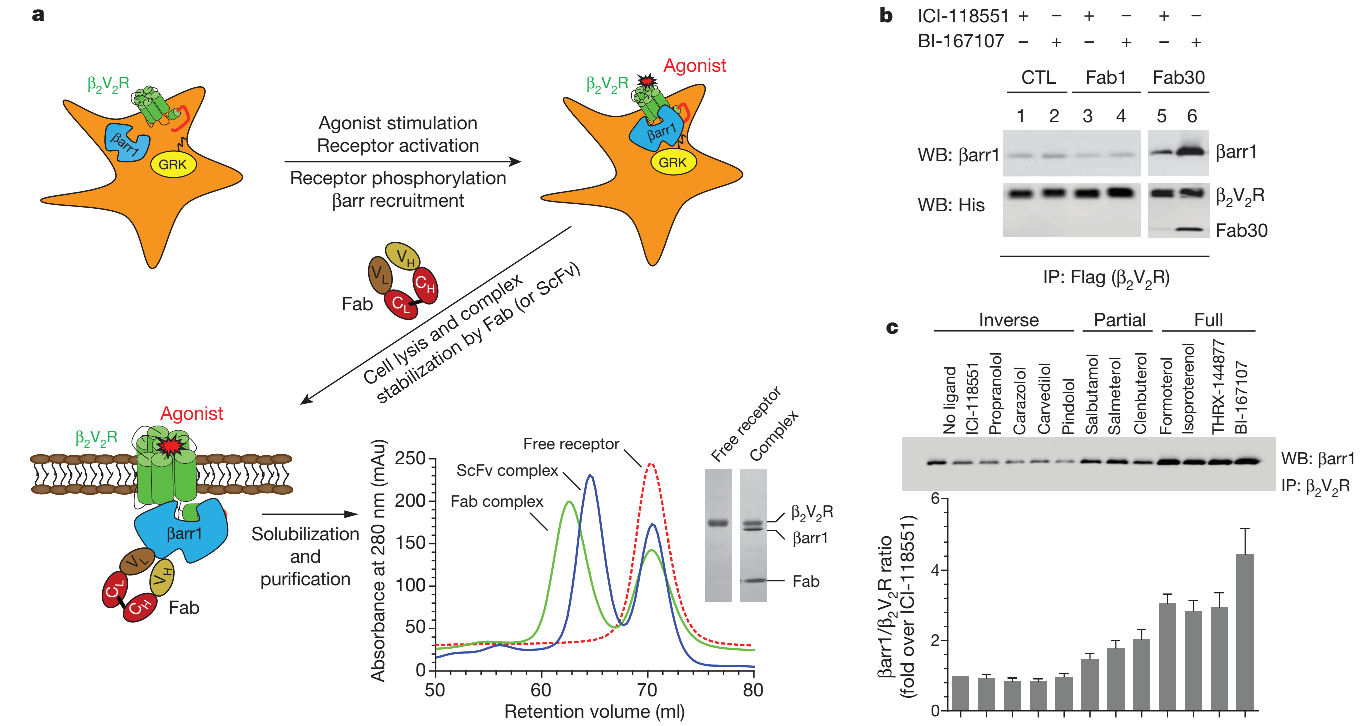


Figure 1 | Formation and functional characterization of a stable agonist–β_2V_2R–β-arrestin-1 signalling complex. **a**, Schematic flowchart of a novel purification strategy to isolate β_2V_2R–β-arrestin-1–Fab30 complex and large-scale production and separation of agonist–β_2V_2R–β-arrestin-1–Fab30/ScFv30 complex from the free receptor by size exclusion chromatography (Superdex 200, 16/600 prep grade). The T4L domain is attached at the N terminus of the β_2AR. βarr, β-arrestin. C_H, constant domain of heavy chain; C_L, constant domain of light chain; V_H, variable domain of heavy chain; V_L, variable domain of light chain. **b**, Isolation of β_2V_2R–β-arrestin-1 complex requires Fab30 and is agonist dependent. Cells were stimulated either with inverse agonist (ICI-118551) or agonist (BI-167107) followed by incubation with or without Fab and subsequent purification on Flag M1 beads. CTL, control; IP, immunoprecipitation; WB, western blot. **c**, Formation of β_2V_2R–β-arrestin-1–Fab30 complex follows ligand efficacy. Formation of the complex in response to inverse agonists, partial agonists and full agonists is shown. The data are representative of seven independent experiments. Error bars, s.e.m.

phosphopeptide (V_2Rpp)– β-arrestin-1–Fab30 complex as a reference to extract specific information about the core interaction between the receptor and β-arrestin 1.

We observed a reduction in the HDX rate in the three major loops—the finger loop (55%), the middle loop (16%) and the lariat loop (23%)—of β-arrestin 1 when we compared the HDX-MS profile of the β_2V_2R–β-arrestin-1–Fab30 complex with that of the V_2Rpp–β-arrestin-1–Fab30 complex (Fig. 2 and Extended Data Fig. 3a). Thus, these regions, and especially the finger loop, are likely to be buried (or have reduced solvent exposure) in the β_2V_2R–β-arrestin-1–Fab30 complex, probably through an intricate engagement with the transmembrane receptor core. This finding is consistent with previous electron paramagnetic resonance (EPR) studies on rhodopsin–arrestin interactions, which revealed a crucial involvement of the finger loop of arrestin with the core of rhodopsin[17,20–22]. Interestingly, several regions in both the N and the C domains of β-arrestin 1, in contrast, reveal enhanced HDX rates, indicating that they become more dynamic upon interaction of β-arrestin with the agonist-bound phosphorylated receptor. This observation suggests that the core interaction between β-arrestin 1 and β_2V_2R probably has long-range effects on β-arrestin 1 structure. Previous studies mapping interactions between GPCRs and arrestins suggested that receptors may also interact with the broad concave surfaces of the N and C domains of arrestins[21,23–25]. However, peptides representing these surfaces are not fully represented in our HDX-MS studies, thus limiting our ability to detect these interactions. We also note that our previously published high-affinity agonist radioligand binding data on the T4 lysozyme (T4L)–β_2V_2R–β-arrestin-1–Fab30 complex in membranes, which provides a readout of the fully engaged β-arrestin conformation, suggested that approximately 32% of the receptor is in a high-affinity agonist binding state[13]. This indicates that our HDX-MS data represent an average of two mixed complex populations, one with fully engaged β-arrestin 1 with the receptor and the other displaying partially engaged β-arrestin 1.

Our previous crystal structure of V_2Rpp bound to activated β-arrestin 1 revealed a marked repositioning of the finger loop compared to when it is bound to the inactive β-arrestin 1, presumably because it is primed to engage with the transmembrane core of the activated receptor[13]. To test this we carried out MS-based mapping of the T4L–β_2V_2R–β-arrestin-1 interface using the homobifunctional, primary amine reactive chemical crosslinker disuccinimidyl adipate (DSA). We found that Lys 77 on β-arrestin 1 (towards the distal end of the finger loop) crosslinks with Lys 235 in the third intracellular loop of the β_2AR (Extended Data Fig. 3b–e). These findings are in line with previously published biochemical and biophysical data suggesting an intricate interaction of the receptor core and the finger loop in arrestins. As an additional control for the close proximity of these residues, we created a series of mutants with single cysteine substitutions around Lys 235 in the N-terminal end of the third intracellular loop of the β_2V_2R (amino acids 231–236) and in the finger loop around Lys 77 of β-arrestin 1 (amino acids 75–79) and evaluated the formation of disulphide-trapped complexes in pairs of receptor and β-arrestin-1 mutants. Consistent with our chemical crosslinking data, cysteines engineered at position 235 of the receptor and at position 78 in β-arrestin 1 yielded the most robust disulphide-trapped complex, suggesting a close proximity of these two residues in the complex (Extended Data Fig. 4). Taken together these findings demonstrate a direct interaction of the finger loop with the receptor core.

We next employed single-particle electron microscopy (EM) to examine the architecture and conformational dynamics of β_2V_2R–β-arrestin-1 complexes. Owing to the asymmetric nature and small size of these complexes (~150 kilodalton (kDa) and ~125 kDa for the Fab and ScFv complexes, respectively) characterization attempts with cryo-EM were

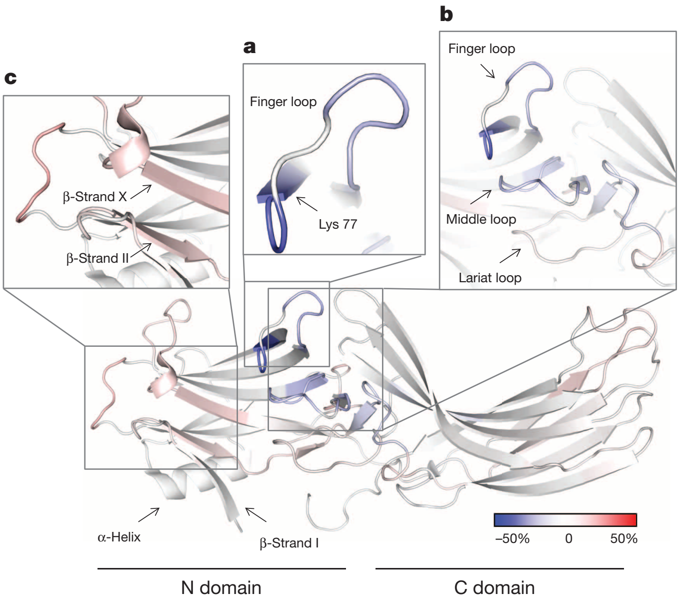


Figure 2 | HDX-MS analysis reveals potential interface between β_2V_2R and β-arrestin 1. a–c, Differential HDX rates of β-arrestin 1 in the β_2V_2R–β-arrestin-1–Fab30 versus V_2Rpp–β-arrestin-1–Fab30 complexes were mapped onto the β-arrestin-1 crystal structure (Protein Data Bank (PDB) accession 4JQI). Blue and red colour coding indicate the β-arrestin-1 regions that exchange slower and faster, respectively, in the β_2V_2R–β-arrestin-1–Fab30 complex when compared to the V_2Rpp–β-arrestin-1–Fab30 complex. Boxed regions with significant HDX rate changes are enlarged in **a**–**c**. The HDX rates of the finger loop (residues 63–75) (**a**), middle loop (residues 129–140) and lariat loop (residues 274–300) (**b**) became slower, whereas those of other regions, for example, β-strand I, II and X in the N domain (**c**) became faster in the β_2V_2R–β-arrestin-1–Fab30 complex when compared to the V_2Rpp–β-arrestin-1–Fab30 complex.

not successful and we thus applied negative-stain EM, which provides adequate contrast for alignment of small particle projections. This approach also enabled a direct comparison with our earlier negative-stain EM analysis of the β_2AR–Gαs protein complex[9]. As in that work, here we used a T4L fusion at the N terminus of the receptor (referred to as T4L–β_2V_2R) to provide a marker for the receptor orientation[9]. The negative-stain EM visualization showed a monodisperse particle population (Fig. 3a and Extended Data Fig. 5) and we applied reference-free alignment and classification to obtain two-dimensional averages of the complex.

The majority of averages of the β_2V_2R–β-arrestin-1–Fab30 complex revealed distinct projection profiles of an ovoid density, attributed to the receptor in partially flattened detergent micelle, with an attached T-like density attributed to the Fab30–β-arrestin-1 complex (Fig. 3b and Extended Data Fig. 6a). Comparisons with averages of the β_2V_2R–β-arrestin-1–ScFv30 complex identify the Fab30 density engaging the middle of β-arrestin 1, in agreement with our recent crystal structure of β-arrestin-1–Fab30 co-crystallized with the V_2Rpp (Fig. 3b and Extended Data Fig. 6b). In this conformation β-arrestin 1 appears to hang off the receptor via a single point interaction presumably involving only the flexible V_2Rpp fused on β_2AR. The flexible nature of this interaction is further supported by the variable receptor orientation in these averages, as judged by the T4L domain positioning. It is possible that this 'hanging' arrestin conformation based on the V_2Rpp–β-arrestin-1 interaction represents a transient intermediate step in the recruitment process that has been stabilized by Fab30. Strikingly, we also observe a substantial number of class averages, representing ~37% of particles, in which β-arrestin 1 forms a much more extensive interface with the receptor, employing roughly the opposite face of the Fab30 binding region (Fig. 3b, bottom). The observed fraction of particles displaying the extensive interface is in agreement with our previous radioligand binding results on the T4L–β_2V_2R–β-arrestin-1–Fab30 complex in membranes, which suggested that approximately 32% of the receptor is in a high-affinity agonist binding state[13]. This observation also raised the possibility that β-arrestin 1 fully engages the receptor through a second set of weak interactions.

To stabilize this weak interaction, we developed an approach whereby the β_2V_2R–β-arrestin-1–Fab30/ScFv30 complex is crosslinked by exposure to a glutaraldehyde-containing buffer zone while migrating through a size exclusion column (Extended Data Fig. 7a). This method facilitated near complete crosslinking of preformed complexes at relatively high concentrations and simultaneously enabled the isolation of highly monodisperse sample (Extended Data Figs 7b, c, 8, 9).

EM classification and averaging of the crosslinked β_2V_2R–β-arrestin-1–Fab30/ScFv30 complexes revealed distinct views of a uniform particle architecture, suggesting that crosslinking stabilized a single complex conformer (Fig. 3c). More importantly, the averages show that arrestin interacts extensively with the receptor in a configuration that appears very similar to the one observed in the smaller fraction (~37%) of the

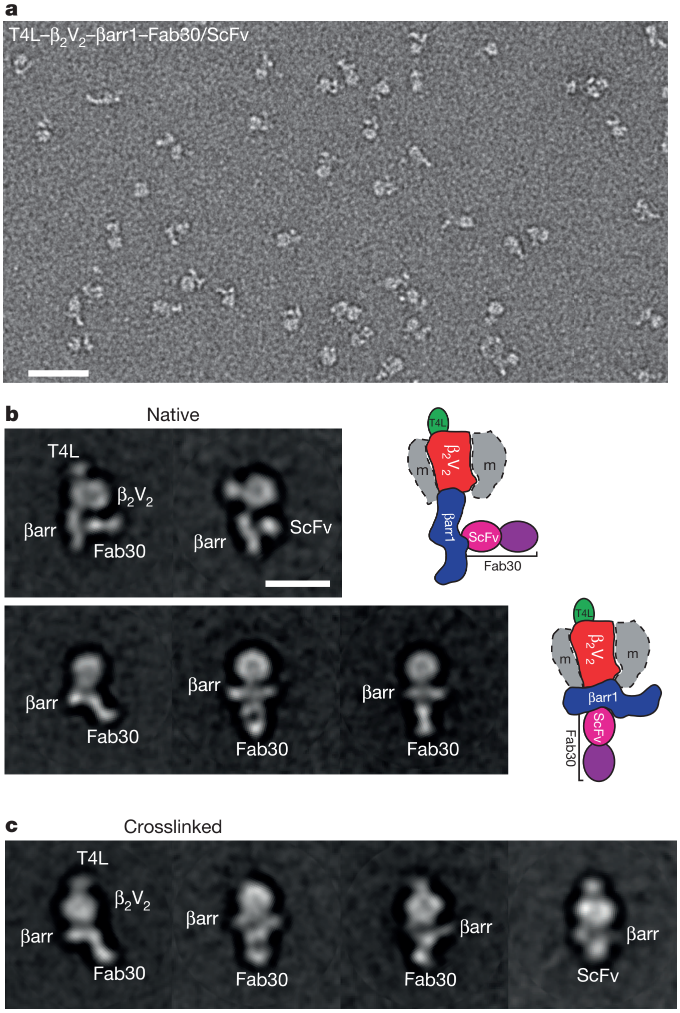


Figure 3 | Single-particle EM analysis of the β_2V_2R–β-arrestin-1–Fab30/ScFv30 complex. a, Representative raw EM image of negative-stained T4L–β_2V_2R–β-arrestin-1–Fab30/ScFv30 complexes. βarr, β-arrestin. Scale bar, 25 nm. **b**, Representative class averages of the native T4L–β_2V_2R–β-arrestin-1–Fab30/ScFv30 complex. Class averages of particles displaying the loose 'hanging' interaction (top) and the fully engaged 'tight' interaction (bottom) are presented. m, LMNG detergent micelle. Scale bar, 10 nm. **c**, Representative class averages of the 'on-column' crosslinked T4L–β_2V_2R–β-arrestin-1–Fab30/ScFv30 complex. Upon crosslinking, the majority of class averages display the tight (fully engaged) β-arrestin-1 conformation, similar to a fraction (~37%) of particles observed in the non-crosslinked complex.

native complex. The conformational stabilizing action of the crosslinking is also evidenced by the consistent position of the T4L projection profile, in contrast to the variable positioning observed in averages of the native complex. To better characterize the β_2V_2R–β-arrestin-1 assembly, we employed the random conical-tilt approach[26] to calculate low-resolution three-dimensional maps (~29 Å) from selected classes of the crosslinked complex (Extended Data Fig. 10). The three-dimensional reconstructions show distinct densities for the main complex components, in full agreement with our domain assignment in the two-dimensional projections averages (Fig. 4a and Extended Data Fig. 10). The receptor-containing region appears ovoid due to the large micelle 'belt' characteristic of the lauryl maltose neopentyl glycol (LMNG) detergent, as we also observed in the case of the β_2AR–Gαs complex[9]. A protrusion on one end of the receptor–micelle globular density represents the T4L domain that marks the receptor extracellular region. On the opposite side, the β-arrestin 1 density lies longitudinally on the receptor, engaging roughly the opposite side of the Fab30-interacting region. In this configuration, both β-arrestin domains appear to engage the receptor but one of the domains lies mostly outside the interacting zone.

The HDX-MS, chemical crosslinking and disulphide trapping data allowed us to constrain the modelling of the T4L–β_2AR and β-arrestin-1–Fab30 crystal structures within the density of the EM three-dimensional maps and generate a low-resolution model for the overall conformation of the β_2AR–β-arrestin-1 complex (Fig. 4b). This model can accommodate limited rotations and translations of the individual crystal structures, which are also expected to undergo conformational changes upon complex formation. Lys 77 of β-arrestin 1 in our model is placed in close proximity to β_2AR Lys 235, which is located at the end of a helical extension of transmembrane (TM)5 in the β_2AR–$G\alpha_s$ complex[10]. This prompted us to use this structure to model the β_2AR–β-arrestin-1 complex. In our model, β-arrestin 1 forms an extensive interface with the receptor through its N-terminal domain, which includes interactions with the phosphorylated receptor tail and the insertion of the finger loop directly in the receptor core, involving the space between TM3, 5 and 6. We note that the finger loop insertion is probably associated with outward shifts in the positioning of TM helices 3, 5 and 6 and also helix 8. The middle and lariat loops of β-arrestin 1 do not participate in major interactions but reside close to the interface, as suggested by the modest reduction in their HDX rates observed by HDX-MS (Fig. 2c). The relative positioning of these loops is also in agreement with previous EPR studies on visual arrestin in complex with activated and phosphorylated rhodopsin[20,21].

In regards to β_2AR, TM5 and the third intracellular loop in this model locate above the concave β-sheet region of the N-terminal domain of

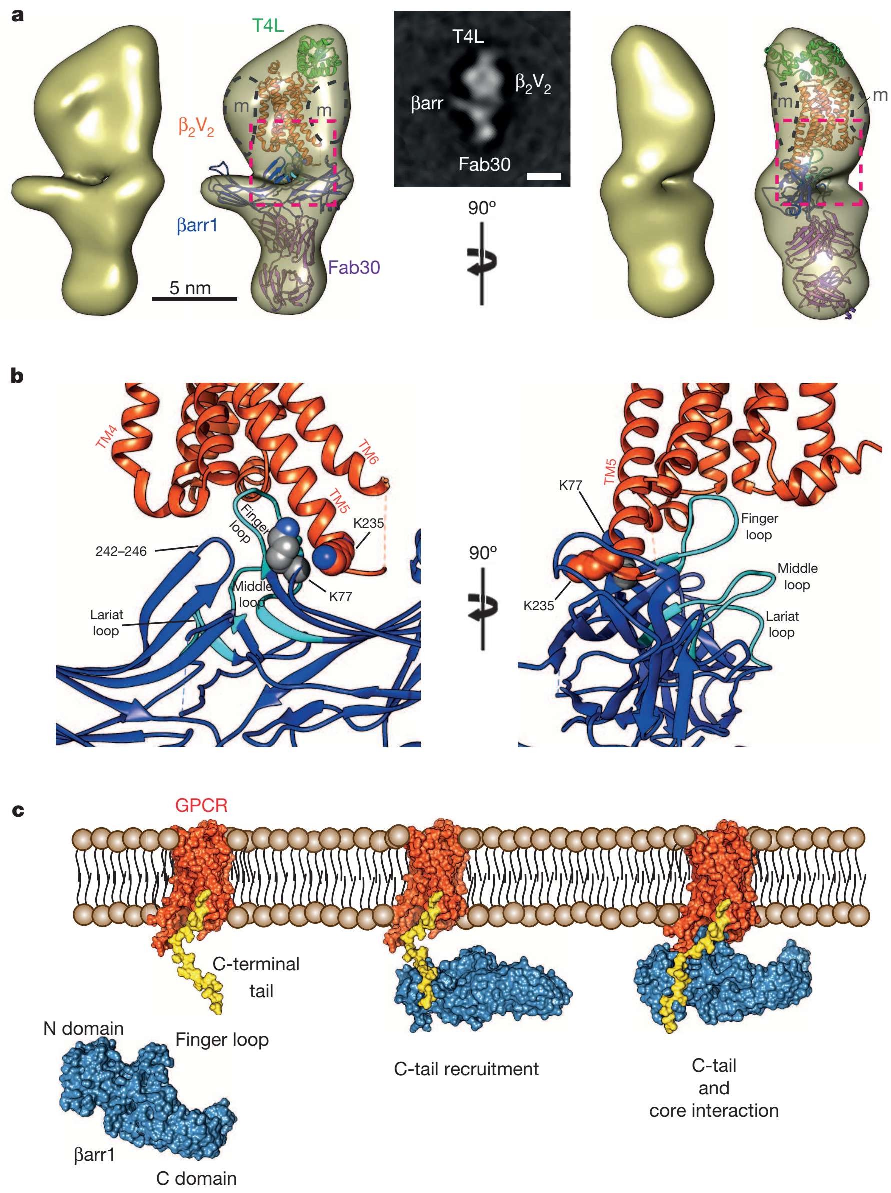


Figure 4 | Structural model of the β_2V_2R–β-arrestin-1–Fab30 complex. a, Views of the T4L–β_2V_2R–β-arrestin-1–Fab30 complex three-dimensional reconstruction with modelled T4L–β_2AR (green–orange; PDB accession 3SN6), β-arrestin-1 (blue; PDB accession 4JQI), and Fab30 (purple; PDB accession 4JQI) crystal structures. The density surrounding β_2V_2R represents the LMNG detergent micelle and is marked by 'm'. βarr, β-arrestin. Scale bar, 5 nm. **b**, Views of the β_2V_2R–β-arrestin-1 interface within the dashed line square of **a**. Areas of β-arrestin 1 with reduced HDX are shown in cyan. Crosslinked Lys 235 of β_2V_2R and Lys 77 of β-arrestin 1 are highlighted. **c**, Illustration of the two-step GPCR–β-arrestin-1 interaction using surface representations of the structures of β_2AR (orange), the phosphorylated C-terminal tail of V_2R (yellow) and β-arrestin 1 (blue). The C-terminal portion of the V_2R peptide (Glu 355–Asp 367) in the right model is positioned as found in the β-arrestin-1–Fab30–V_2Rpp structure (PDB accession 4JQI), whereas the N-terminal portion (Ala 342–Pro 352) was remodelled to connect to the β_2AR C terminus.

β-arrestin 1. The placement of these receptor elements implies that the N terminus of V_2Rpp cannot be in the position observed in the crystal structure of V_2Rpp–β-arrestin-1–Fab30 (ref. 13), suggesting that the V_2R C terminus in the β_2V_2R chimaeric receptor is mobile and repositions itself markedly upon β-arrestin-1 interaction with the receptor core. In contrast to the N-terminal domain, the C-terminal domain of β-arrestin 1 lies mostly outside the interaction zone, apart from the loop of residues 242–246 that is at interacting distance from the short α-helical segment connecting TM3 and TM4 of β_2V_2R. This observation is intriguing considering that mutation of the residues distal to the DRY motif (at the end of TM3) have been reported to directly affect β-arrestin recruitment for a number of GPCRs including the β_2AR[27].

Our results suggest that arrestin probably employs a biphasic mechanism to engage the receptor (Fig. 4c). The first phase involves an interaction between the phosphorylated C-terminal tail of the receptor and the N-terminal domain of arrestin. Given the flexibility and the length of the C-terminal receptor tail, it is expected to act like a fishing line, sampling a wide interaction space at a high rate. The second point of interaction appears weak and involves primarily the insertion of the finger loop within the receptor core, resulting in a longitudinal arrangement of arrestin on the receptor (Fig. 4a, c). This arrangement would most certainly preclude GPCR engagement of G-protein heterotrimers, thereby blocking classical GPCR signalling and inducing desensitization. While it is not yet clear whether the single point interaction resulting in a hanging arrestin configuration has other physiological functions, it seems possible that these might involve recruitment and complex formation with components of the receptor endocytosis and signalling machinery such as clathrin and Gβγ.

METHODS SUMMARY

β_2V_2R, β-arrestin 1 and GRK2^{CAAX} were co-expressed in Sf9 cells. Sixty-six hours post-infection, cells were stimulated with the high-affinity agonist BI-167107 for 30 min at 37 °C. Cells were harvested and lysed by douncing, followed by incubation with purified Fab30. One hour post-incubation, cells were solubilized and purified on a Flag M1 affinity column followed by size exclusion chromatography. The purified complex was subjected to HDX-MS analysis by incubating it with D_2O for various time points followed by pepsin digestion and liquid chromatography (LC)/MS-based identification of peptides. Purified T4L–β_2V_2R–β-arrestin-1–Fab30/ScFv30 complex was embedded in negative stain and visualized by EM. EM two-dimensional averages of the complexes were obtained by ISAC[28] and three-dimensional reconstructions were obtained through the random conical-tilt method[26].

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 13 January; accepted 30 April 2014.

Published online 22 June 2014.

1. Pierce, K. L. & Lefkowitz, R. J. Classical and new roles of β-arrestins in the regulation of G-protein-coupled receptors. *Nature Rev. Neurosci.* **2,** 727–733 (2001).
2. Shukla, A. K., Xiao, K. & Lefkowitz, R. J. Emerging paradigms of β-arrestin-dependent seven transmembrane receptor signaling. *Trends Biochem. Sci.* **36,** 457–469 (2011).
3. Lefkowitz, R. J. & Shenoy, S. K. Transduction of receptor signals by β-arrestins. *Science* **308,** 512–517 (2005).
4. Pierce, K. L., Premont, R. T. & Lefkowitz, R. J. Seven-transmembrane receptors. *Nature Rev. Mol. Cell Biol.* **3,** 639–650 (2002).
5. DeWire, S. M., Ahn, S., Lefkowitz, R. J. & Shenoy, S. K. β-Arrestins and cell signaling. *Annu. Rev. Physiol.* **69,** 483–510 (2007).
6. Rasmussen, S. G. *et al.* Crystal structure of the β_2 adrenergic receptor–Gs protein complex. *Nature* **477,** 549–555 (2011).
7. Weis, W. I. & Kobilka, B. K. Structural insights into G-protein-coupled receptor activation. *Curr. Opin. Struct. Biol.* **18,** 734–740 (2008).
8. Rosenbaum, D. M., Rasmussen, S. G. & Kobilka, B. K. The structure and function of G-protein-coupled receptors. *Nature* **459,** 356–363 (2009).
9. Westfield, G. H. *et al.* Structural flexibility of the Gαs α-helical domain in the β_2-adrenoceptor Gs complex. *Proc. Natl Acad. Sci. USA* **108,** 16086–16091 (2011).
10. Rasmussen, S. G. *et al.* Crystal structure of the human β_2 adrenergic G-protein-coupled receptor. *Nature* **450,** 383–387 (2007).
11. Rasmussen, S. G. *et al.* Structure of a nanobody-stabilized active state of the β_2 adrenoceptor. *Nature* **469,** 175–180 (2011).
12. Oakley, R. H., Laporte, S. A., Holt, J. A., Caron, M. G. & Barak, L. S. Differential affinities of visual arrestin, β arrestin1, and β arrestin2 for G protein-coupled receptors delineate two major classes of receptors. *J. Biol. Chem.* **275,** 17201–17210 (2000).
13. Shukla, A. K. *et al.* Structure of active β-arrestin-1 bound to a G-protein-coupled receptor phosphopeptide. *Nature* **497,** 137–141 (2013).
14. Goodman, O. B. Jr *et al.* β-Arrestin acts as a clathrin adaptor in endocytosis of the β_2-adrenergic receptor. *Nature* **383,** 447–450 (1996).
15. Nobles, K. N., Guan, Z., Xiao, K., Oas, T. G. & Lefkowitz, R. J. The active conformation of β-arrestin1: direct evidence for the phosphate sensor in the N-domain and conformational differences in the active states of β-arrestins1 and -2. *J. Biol. Chem.* **282,** 21370–21381 (2007).
16. Xiao, K., Shenoy, S. K., Nobles, K. & Lefkowitz, R. J. Activation-dependent conformational changes in β-arrestin 2. *J. Biol. Chem.* **279,** 55744–55753 (2004).
17. Gurevich, V. V. & Gurevich, E. V. The molecular acrobatics of arrestin activation. *Trends Pharmacol. Sci.* **25,** 105–111 (2004).
18. Chung, K. Y. *et al.* Conformational changes in the G protein Gs induced by the β_2 adrenergic receptor. *Nature* **477,** 611–615 (2011).
19. Konermann, L., Pan, J. & Liu, Y. H. Hydrogen exchange mass spectrometry for studying protein structure and dynamics. *Chem. Soc. Rev.* **40,** 1224–1234 (2011).
20. Kim, M. *et al.* Conformation of receptor-bound visual arrestin. *Proc. Natl Acad. Sci. USA* **109,** 18407–18412 (2012).
21. Hanson, S. M. *et al.* Differential interaction of spin-labeled arrestin with inactive and active phosphorhodopsin. *Proc. Natl Acad. Sci. USA* **103,** 4900–4905 (2006).
22. Zhuang, T. *et al.* Involvement of distinct arrestin-1 elements in binding to different functional forms of rhodopsin. *Proc. Natl Acad. Sci. USA* **110,** 942–947 (2013).
23. Gimenez, L. E., Vishnivetskiy, S. A., Baameur, F. & Gurevich, V. V. Manipulation of very few receptor discriminator residues greatly enhances receptor specificity of non-visual arrestins. *J. Biol. Chem.* **287,** 29495–29505 (2012).
24. Gurevich, V. V. & Gurevich, E. V. Structural determinants of arrestin functions. *Prog. Mol. Biol. Transl. Sci.* **118,** 57–92 (2013).
25. Lohse, M. J. & Hoffmann, C. Arrestin interactions with G protein-coupled receptors. *Handb. Exp. Pharmacol.* **219,** 15–56 (2014).
26. Radermacher, M., Wagenknecht, T., Verschoor, A. & Frank, J. Three-dimensional reconstruction from a single-exposure, random conical tilt series applied to the 50S ribosomal subunit of *Escherichia coli*. *J. Microsc.* **146,** 113–136 (1987).
27. Kim, K. M. & Caron, M. G. Complementary roles of the DRY motif and C-terminus tail of GPCRS for G protein coupling and β-arrestin interaction. *Biochem. Biophys. Res. Commun.* **366,** 42–47 (2008).
28. Yang, Z., Fang, J., Chittuluru, J., Asturias, F. J. & Penczek, P. A. Iterative stable alignment and clustering of 2D transmission electron microscope images. *Structure* **20,** 237–247 (2012).

Acknowledgements We thank D. Capel for technical assistance, V. Ronk, D. Addison and Q. Lennon for administrative support, R. K. Sunahara for stimulating discussions and Alex R. B. Thomsen for critical reading of the manuscript. We acknowledge support from the National Institutes of Health Grants DK090165 (G.S.), NS028471 (B.K.K.), GM072688 and GM087519 (A.A.K. and S.K.), HL075443 (K.X.), HL16037 and HL70631 (R.J.L.), from the Mathers Foundation (B.K.K.), GM60635 (P.A.P.) and from the Pew Scholars Program in Biomedical Sciences (G.S.). R.H. and S.S.S. were supported by a research grant from the Canadian Institutes of Health Research (MOP-93725). R.I.R is supported by a postdoctoral fellowship from Coordenação de Aperfeiçoamento de Pessoal de Nível Superior. R.J.L. is an investigator with the Howard Hughes Medical Institute.

Author Contributions A.K.S. designed and optimized procedures for forming and purifying the complex, executed and optimized the on-column crosslinking protocol and provided the preparations of complex used for EM, HDX-MS and crosslink mapping experiments with assistance from P.T.-S. R.I.R. and L.-Y.H. performed biochemical and pharmacological characterization of the complex. G.H.W. performed EM analysis assisted by M.S., A.N.O. and A.M.D. and supervised by G.S. K.X. performed the HDX-MS experiments assisted by S.L., J.Q., A.W.K. and A.B., performed the crosslink mapping experiments assisted by J.Q. and A.W.K., and designed the disulphide trapping experiments carried out by M.C. V.L.W. Jr supervised the initial phase of the HDX-MS experiments. C.-R.L., L.-L.G., J.-M.S. and X.C. synthesized the high-affinity agonist BI-167107. R.H. and S.S.S. provided the linker sequence, vector and advice on ScFv conversion and expression. X.J.Y. and B.U.K. contributed in assessing various methods of complex formation. P.A.P. provided advice on implementation of ISAC[28]. S.K. and A.A.K. provided the phage display library and protocols for Fab selection, expression and purification. B.K.K. conceived the on-column crosslinking strategy, advised A.K.S. on its execution and optimization, assisted with molecular modelling of the complex and participated in supervision of the project. G.S. directly supervised the EM studies, performed the molecular modelling of the complex and supervised overall project execution. R.J.L. supervised overall project design and execution. A.K.S., G.H.W., K.X., G.S., B.K.K. and R.J.L. participated in data analysis and interpretation. A.K.S., G.H.W., K.X., G.S., B.K.K. and R.J.L. wrote the manuscript. All authors have seen and commented on the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to R.J.L. (lefko001@receptor-biol.duke.edu), B.K.K. (kobilka@stanford.edu) or G.S. (skinioti@umich.edu).

@NATUREJOBS Follow us on Twitter for the latest on science jobs twitter.com/naturejobs

NATUREJOBS FACEBOOK Science-careers advice & tips www.facebook.com/naturejobs

NATUREJOBS For the latest career listings and advice www.naturejobs.com

ART4ALL/SHUTTERSTOCK

MENTAL HEALTH

Stressed students reach out for help

Graduate students struggling with the stresses of their work and lives can tap into multiple avenues of support.

BY JULIE GOULD

Sarah Gossan got mostly 'A's during her undergraduate astrophysics programme at Cardiff University, UK, and graduated at the top of her class in 2012. In her third year, she started to study for the Graduate Record Examination (GRE), a standard test for admission to US graduate programmes, in the hope of starting a PhD in gravitational waves at the California Institute of Technology (Caltech) in Pasadena after she graduated.

What her undergraduate peers and supervisors did not know was that she was struggling to deal with severe bulimia and depression. "The stress from research towards the end of my third year and the grad-school application process led to a relapse for about 8 months," Gossan says. "I almost took a sick sense of pride in performing well academically while mentally ill."

Gossan did not reach out for help. "I was too embarrassed to ask for help for the bulimia, and it continued to get worse," she says. Then, once she was accepted at Caltech, the transition to the new environment caused her yet more stress. But she kept quiet because she did not want to appear weak, particularly as a woman in science. "I was afraid of being painted as 'just another emotional woman,'" she says. The consequences were dismal: she failed the PhD qualifying exams twice in addition to an exam on classical physics, went on strong medication and did not attend classes for almost five months.

Gossan's experience is not unique. Maintaining mental health as a researcher-in-training can be a contradiction in terms. Many doctoral students are free to pursue a scientific field of their choice and, at least in theory, get an opportunity to become a leading researcher in that field. But the need to publish often, conduct research independently, constantly apply for funding and meet the needs of supervisors can create substantial emotional and mental strain, anxiety and pressure. These hurdles can adversely affect a PhD student's emotional wellbeing, especially if they are not expecting them — or do not know how to surmount them.

There is a high risk of developing a major psychiatric illness, such as depression, schizophrenia, and bipolar or anxiety disorder, between the ages of 18 and 24 — just when students are pursuing degrees, says Victor Schwartz, who is medical director at the Jed Foundation, a charitable organization in New York that aims to reduce suicide rates and improve mental health for university students. "This is a time of transition from adolescence into adulthood, and often from undergraduate to graduate studies," he explains. "Students experience many firsts, including new lifestyles, friends, roommates, cultures and ways of thinking." Graduate students move off campus and become further removed from support networks, conduct more independent research and face uncertain career prospects, thanks to an unsteady regional and global job market, he says. Combine academic stresses with this transition, and it is not surprising that many doctoral students struggle to maintain mental health.

Nearly one-fifth of the general US populace over the age of 18 — and 13% of master's ▶

▶ or PhD students — suffer from anxiety or depression (D. Eisenberg *et al. Am. J. Orthopsychiatry* **77,** 534–542; 2007). Those in doctoral programmes are especially susceptible, says Catherine McAteer, head of student services at University College London (UCL). "PhD students tend to spend a lot of time by themselves doing their research in a lab or writing their theses, and isolation is often an instant pathway to depression and anxiety." To combat the procrastination that goes hand-in-hand with isolation, UCL offers classes for PhD students to help them to focus attention on the present (see 'Mind tricks').

In a 2013 US poll of 41,847 undergraduate and graduate students, almost one-third said that they had "felt so depressed that it was difficult to function" in the past year. And nearly half said that their academic programme — their studies, research, lab colleagues and supervisors — had been "very difficult to handle" in the past year.

Schwartz's advice to people who are struggling with depression, anxiety and other disorders is to reach out to others — whether they are friends, loved ones or counselling services. Schwartz and McAteer both advocate taking advantage of on-campus mental-health services, which offer options such as group counselling, one-on-one sessions and peer support, in which students form a network that aims to promote mental health and provide confidential help. Campus peer groups are becoming more common: examples include Active Minds, which is based in Washington DC but has groups around the world, Peer Ears at the Massachusetts Institute of Technology in Cambridge and Cause for Concern at UCL.

MIKE MCGREGOR/ACTIVE MINDS INC.

Active Minds runs groups, like this one in Pennsylvania, that encourage students to discuss depression.

MIND TRICKS

How mindfulness works

Mindfulness is a therapeutic practice that helps to increase awareness of the present, which can improve thinking habits and mental health. Imagine, for instance, that your experiment has gone wrong: the data are not coming together and your deadline is tomorrow. Your usual response is to panic. Instead, you can:

- Take three deep breaths. This stimulates the vagus nerve, which releases a chemical called acetylcholine that will calm you down.
- Concentrate on the here and now in a non-judgemental way. Rather than blaming yourself, take a step back. By objectively acknowledging your frustrations, you will be able to see the problems more clearly and focus on how to solve them.
- Keep your focus on a single object, idea or sensation rather than letting your mind wander off.
- Stay aware of your body and its response to internal and external stimuli.
- Reframe your emotions in a positive way. In the wake of negative thoughts or experiences, this can help you to react less emotionally and to be more resilient.
- Try to be as objective as possible in terms of the way you think about yourself. **J.G.**

To reduce the likelihood of developing mental-health problems, doctoral researchers should try to build a solid and trustworthy peer group in the early days of their programme, says Charlotte Vaughan, the disability adviser for mental health at UCL. This can be accomplished by joining discipline-based societies and clubs, or networks set up by the university mental-health services. "Most of all, we want to make sure that students are aware of the possible mental-health conditions they may face," she says, "and know where they need to go if they think they're running into trouble."

It was not until earlier this year that Gossan started looking for serious help, after her partner asked her to do so. She began by speaking to Christian Ott, her supervisor at Caltech, who reassured her that "many, if not most people, cope with a whole range of mental issues". Ott had dealt with his own problems in the past, and he values being open about the topic of mental health. "I made it clear that it is a common thing to run into such problems and that getting help and looking ahead is the important thing to do," he says.

Once she knew how Ott felt, Gossan spoke to him whenever she found she was falling back into old, unhealthy habits. "I work very actively to prevent a relapse," says Ott. "This sometimes involves telling her specifically what to do or not to do."

If left untreated, mental-health problems can lead to suicide, says Charles Reynolds, a behavioural and community-health scientist at the Graduate School of Public Health of the University of Pittsburgh in Pennsylvania. A US survey in 2009 found that 4% of graduate students had "seriously considered attempting suicide" in the past 12 months (D. J. Drum *et al. Prof. Psychiatry Res. Prac.* **40,** 213–222; 2009), and in 2011, the American College Health Association reported that suicides were the leading cause of death in undergraduate and graduate students. "We need to remove this stigma attached to mental-health problems and find a way to get students to talk," Reynolds says.

Active Minds and Peer Ears are helpful in terms of intervention and treatment, say those involved in the networks. Talking discreetly with peer-support-group representatives can help to ease the fear that opening up about mental-health woes will have academic and other ramifications. Indeed, Gossan was warned by a fellow doctoral student against telling a supervisor about her depression. The colleague, who had suffered from mental-health disorders, told her that people would view her as unreliable and would not want to work with her, illustrating that advice from untrained peers may not be always reliable.

"We need to remove this stigma attached to mental-health problems and find a way to get students to talk."
Charles Reynolds

Ultimately, experts say, many mental-health issues, including depression, can be resolved only by talking to others — whether a counsellor, supervisor or peer representative. Gossan is grateful that Ott has been there for her. "He helped me through many anxiety attacks," she says, "and without his support I think I probably would have dropped out by now." ■

Julie Gould *is editor of Naturejobs online.*

THE DEATH OF IMMORTALITY

Life lessons.

BY KYLE L. WILSON & ANDREW B. BARBOUR

“In all my years, I’ve never …” the doctor trembled while checking the readings. “Your son, he’s going to die.”

Tears streamed down the mother’s face. She gripped her newborn tightly. “What does that mean, he’s going to *die*? How, how long does that take?”

The doctor hesitated. “Maybe a century. The genetic implants did not take. Somehow, the testing failed.” Pulling up a millennium-old document in the Venter Laboratories database, he found the passage. “Continually shrinking telomeres. He will age and develop archaic tumours and, eventually, cancer. His body will fail. We haven’t seen anything like this since the Mortal Era.”

Jaw clenched, she chewed on her words. “Do you know how many centuries it took to get approval for a child? And now what … he’s like … like one of our dogs!”

Rosalind had followed the child since he first left his arcology six months ago. For much of Michael’s journey, they were not alone. The world watched, captivated by the mortal’s decision to leave his risk-free home. Rosalind, motivated by desire for advancement in a stagnant workforce, filmed his trek.

“Michael grew a grey hair. He’s so different! The fans will go crazy when they see. And our updated ‘road map,’ as he terms it, shows that we will lose communication for weeks. He wants to climb some mountain in a place called the Himalayas. And after that, Europe!”

Her boss was baffled. “You mean Europa? Surely he’d want to visit the moons …”

“No, Europe. Sir, I don’t understand him.”

“Who wants to go to that dreadful place? Since the floods it’s been deserted. This story is great Rosy, can you figure out the mindset of the last person that will ever die?”

“He says it’s to connect with his roots, to experience history. But I think he’s insane. This is becoming too much for me. I need to be transferred.”

“That’s impossible, Rosy. No one else will risk leaving the arcologies. We need your reports for our ratings. Keep this up and I’ll promote you within a century.”

Despite her years, the reporter maintained a youthful complexion with rose-red cheeks and blonde hair. Next to her, the grey-haired and sun-baked man. The stunning 400-year-old girl and ancient 90-year-old man traversed the sands of the old-world desert.

He grinned like a mad Cheshire at the sight of the dunes. “The Arabian desert. Where Lawrence fought all those years ago.”

JACEY

She smiled warmly with youthful radiance, comforted by the decades-long bond the pair had formed. “I never knew such a place existed. Michael, it’s stunning!” She was rather hot and needed some water. To her left, off in the distance, she noticed a snake slithering away. *Impossible a thing could survive here.*

They camped under the Milky Way’s glow. She worked up a fire, a skill learned from Michael years ago. *Hundreds of years and I had never made a fire before …*

Suddenly, Michael started to cough. His once-strong legs shook as they strained to hold him upright. “Rose, I’m feeling weak now. The weakest since I left the hospital.” He sounded worried, despite his typical confidence.

She was worried as well. She didn’t know how to act around a dying man, no one did. “Michael, why don’t we head back to the city —”

A crackling voice interrupted. “Where are you?! No updates in three months! We need a new report, we can only show so much old footage before people lose interest.”

She maintained her composure while concerned for her dying friend. “It’s not important now, Sir. I’ve been recording our observations and I’ll submit them when we head back to the city.” Abruptly, she turned off her communicator.

“Michael, let’s go back. It’s been *walk walk walk* for decades. You’re fatigued.” The white lie made her sad. *What else could she say?* Only he knew how to feel — after all, he’d spent years reading antiquated books on religion and death.

Back in the city hospital, Michael slipped away while Rosalind desperately gripped his hand. Really, the whole world held him. His death was broadcast live across the entire system, from Earth to Ganymede. *Just like the doctor said: “Old age.”*

The audience watched his last breath. They didn’t know what to make of it, wondering where he would go. Distressed and feeling that lingering existential dread, the world switched channels.

“Welcome to the Records and Application office, how may we serve you … Oh! Rosalind! I *loved* your work on Michael.”

Rose exchanged pleasantries before asking for the necessary forms. “I would like to apply for a child.”

Held lightly in her hand, the pen danced across the tedious application. Upon review, she heard the administrator chuckle while flipping through her forms.

“A girl huh? That’s great, I’m sure she will be just as adventurous as her mother!” The man read on. “Oh, uh … I see you’ve filled out the liability waiver to decline genetic implants. Funny, that’s the fourth time I’ve seen that this week.” His pen darted a note. “Well, the application is in order, but I’m required by law to advise you against this choice. After all, you know better than anyone the severe disability your child will face.” Rosalind, warmly remembering her last days with Michael, nodded in acknowledgement. ■

Kyle L. Wilson *is a PhD student at the University of Calgary, where he studies ecology and evolutionary biology.* **Andrew B. Barbour** *graduated with a PhD in fisheries from the University of Florida and now works as a research associate in neonatology at the Medical University of South Carolina.*